



UNIVERSITÄT REGENSBURG

# **Evaluating Marketing Allocation and Pricing Rules by Monte-Carlo Simulation**

DISSERTATION ZUR ERLANGUNG EINES  
DOKTORGRADES DER WIRTSCHAFTSWISSENSCHAFT

EINGEREICHT AN DER FAKULTÄT FÜR  
WIRTSCHAFTSWISSENSCHAFTEN  
DER UNIVERSITÄT REGENSBURG

Vorgelegt von  
*Daniel Gahler*  
aus Korbach

Berichterstatter:  
Prof. Dr. Harald HRUSCHKA  
Prof. Dr. Guido SCHRYEN

Tag der Disputation: 22.5.2020



# Contents

<b>1</b>	<b>Introduction</b>	<b>9</b>
1.1	Topical Introduction to the Thesis . . . . .	9
1.2	Preliminaries, Notation and Definitions . . . . .	12
1.3	Acknowledgments . . . . .	16
<b>2</b>	<b>Allocation Procedures</b>	<b>17</b>
2.1	Introduction . . . . .	18
2.2	Decision Problem . . . . .	21
2.3	Simulation . . . . .	23
2.3.1	Sales Response Functions . . . . .	24
2.3.2	Determining Parameters of Sales Units . . . . .	24
2.3.3	Disturbances . . . . .	25
2.4	Developed Algorithm . . . . .	27
2.4.1	Exploration . . . . .	27
2.4.2	Exploitation . . . . .	28
2.5	Rules of Thumb . . . . .	31
2.6	Hypotheses . . . . .	32
2.7	Evaluation of Procedures . . . . .	33
2.7.1	Results for Regression Models with Interactions . . . . .	35
2.8	Conclusion . . . . .	37
2.9	Appendix A: The Algorithm in Pseudocode . . . . .	39
2.10	Appendix B: Additional Regression Results . . . . .	40
2.10.1	Comparison of Nested Models . . . . .	40
2.10.2	F-Tests in the Full Models . . . . .	40
2.10.3	Pairwise interactions . . . . .	41
2.11	Appendix C: Supplementary material on generating functions . . . . .	41
2.11.1	Albers' Original Table . . . . .	41
2.11.2	Visually testing S-shaped functions . . . . .	43
<b>3</b>	<b>Pricing Procedures</b>	<b>45</b>
3.1	Introduction . . . . .	46
3.2	Preparing the Simulation . . . . .	48
3.2.1	Decision Problem . . . . .	48
3.2.2	Pricing Rules . . . . .	49

3.2.3	Parameter restrictions on Price Response Functions . . . . .	52
3.2.4	Functional Forms . . . . .	52
3.2.5	Setting Coefficients . . . . .	53
3.2.6	Determining Optimal Prices . . . . .	53
3.2.7	Evaluating Elasticities . . . . .	54
3.2.8	Restrictions between Rules and Functions . . . . .	54
3.2.9	Error Terms . . . . .	54
3.2.10	Iterations . . . . .	55
3.3	Conducting the Simulation . . . . .	55
3.3.1	Experimental Designs . . . . .	55
3.3.2	Performing Tests . . . . .	55
3.3.3	Recommended Rules . . . . .	56
3.3.4	Unrecommended Rules . . . . .	57
3.3.5	Robustness . . . . .	58
3.4	Conclusion . . . . .	58
3.5	Appendix A: Optimization of Parameters in Rules of Thumb . . . . .	59
3.6	Appendix B: Derivations of Coefficients . . . . .	60
3.7	Appendix C: Derivations of Prices . . . . .	63
3.8	Appendix D: Derivations of Elasticities . . . . .	68
<b>4</b>	<b>Statistical Procedures</b>	<b>69</b>
4.1	Introduction . . . . .	70
4.2	Methodology . . . . .	72
4.2.1	Interaction Coefficients in full $N$ -way ANOVA models . . . . .	72
4.2.2	Derivation of null hypotheses . . . . .	89
4.2.3	Complete Null Hypothesis . . . . .	91
4.3	Application . . . . .	92
4.4	Conclusion . . . . .	95
4.5	Appendix A: Proofs . . . . .	96
4.6	Appendix B: Further Comments . . . . .	102
<b>5</b>	<b>Conclusions</b>	<b>111</b>
	<b>Index</b>	<b>117</b>

# List of Figures

2.1	True Function and Fitted Parabola . . . . .	29
2.2	Means when switching after $t$ periods . . . . .	31
2.3	Box-Plot of Optimality and Sales . . . . .	37
2.4	Undesirable function shapes . . . . .	44



# List of Tables

2.1	Comparison of procedures . . . . .	19
2.2	Function Properties for Parameter Generation . . . . .	25
2.3	Arithmetic Means of Dependent Variables for Each Procedure . . . . .	34
2.4	Main Effect Regression Models . . . . .	35
2.5	Pairwise Interaction Effects Regression Models . . . . .	42
2.6	Original Table from Albers . . . . .	42
3.1	Properties . . . . .	52
3.2	Recommended Rules . . . . .	56
3.3	Unrecommended Rules . . . . .	57
4.1	Table with observed cell values . . . . .	72
4.2	Table of cells to be filled in . . . . .	74
4.3	Filled in Table . . . . .	79
4.4	Table of cells to be filled in . . . . .	84
4.5	Complete table . . . . .	86
4.6	summary(FULL1) . . . . .	93





# Chapter 1

## Introduction

### 1.1 Topical Introduction to the Thesis

Every measurement is prone to error (JCGM-WG1, 2008, 1.2.3.).

Let us demonstrate what we mean by that with the following simple example: When reading the temperature off a mercury-thermometer, several things can lead to the value being incorrect. The thermometer might be placed at a point that is not representative of the areas average temperature, like next to an oven or in the shade. The resolution of the scale is bounded, so temperature differences smaller than 0.1 degrees might be difficult to observe. If the thermometer was produced too cheaply, the scale might even be wrong, if the reader has bad eyesight the measurement will not be precise. Finally, the interpretation might be wrong, e.g. someone interprets a Celsius scale as Fahrenheit. While a digital thermometer can potentially solve all of these issues, it is still possible that two identically manufactured thermometers observe slightly different temperatures at the same time and place. That is because the digital device introduces a whole new set of errors, like rounding errors. In fact, a lot of work goes into gauging thermometers (JCGM-WG1, 2008, H.3.), and minimizing possible errors.

Not only is it important to *realize* that measurements are prone to error, one should also know how to *handle* these errors. As every empirical area of science relies on data that comes from some kind of measurements, everyone working in such areas must know how they occur and how to treat them (Stier, 1999, I.1). In economics, errors appear in several areas. While there is extensive research on different kinds of errors and how to tell them apart, that will not be the topic of this thesis. Instead, the novel approach this thesis brings to the field will be the introduction of an error term in areas where they have not been analyzed yet, or only spuriously.

In this thesis mathematical models will be used to approximate the reality of a market situation. The most useful tool in this thesis will be so called *sales response functions* (Hruschka, 1996, Ch 5.2, Ch 7.3). Fixing a unit of time, say one period being

one month, the sales response function will take a marketing variable (like the price of a product or advertisement budget spent on that product in a given month), and it will return the number of sold items at the end of the month.

$$\text{Sales Response Function : Marketing Variable} \rightarrow \text{Sales} \quad (1.1)$$

Using a simple model like ‘The quantity of sold units is 10 minus the price’ will lead to sales of 5 units at a price of 5, and 7 units at a price of 3.

$$Q : \{\text{Price } p \in \mathbb{R}^+\} \rightarrow \{\text{Sales } Q \in \mathbb{R}^+\} \quad (1.2)$$

$$p \mapsto Q(p) = 10 - p \quad (1.3)$$

This particular sales response function comes with several problems, most importantly, it does not include an error term. Setting a price of 3 will *always* lead to sales of 7 independent of the month, market fluctuations or other circumstances. But knowing that measurements must be prone to error, a better sales response function should include an error term.

So we take a (fair) coin and flip it. When the coin shows ‘heads’, we treat it as a value of  $e = 1$ , when it shows ‘tails’, we treat it as a value of  $e = -1$ . We have no ordinary way of influencing the outcome of the coin flip or knowing the outcome before observing it. For the purpose of this introduction, this seems to be an adequate way to model a simple error. If we add the value of the coin flip to the sales response function, setting a price of 3 might lead to sales of 6 or 8, dependent on what face comes up. We now have a sales response function with an error term:

$$Q_e : \{\text{Price } p \in \mathbb{R}^+\} \rightarrow \{\text{Sales } Q \in \mathbb{R}^+\} \quad (1.4)$$

$$p \mapsto Q_e(p) = 10 - p + e \quad (1.5)$$

We will use sales response functions with error terms to model markets and evaluate the performance of procedures. The performance of those procedures however will not usually be determined by the value of the sales response function. Indeed, a business usually needs to pursue a more complex objective than optimizing sales of a single month. This so called *objective function* will be optimized dependent on the variables that influence it. A common version is sum of all profits of a given planning horizon. If we assume that the variable costs of producing one unit of our product to be 2, then the profit margin will be the price we set minus the costs. A price of 3 will lead to a margin of 1, which will be the profit for each sold item. At a constant price of 3, the sales and therefore the profit will be either 6 or 8, and adding them over 12 periods will lead to a profit between  $6 \cdot 12 = 72$  and  $8 \cdot 12 = 96$ .

However, in reality, we do not know the mechanism behind the sales response function, all we can do is give a price as input and observe an error-loaded sales value as output, which will then be used to determine the objective, i.e. the profit. The sales response function itself, i.e. (1.4) is a black-box. The determination of the best input values with no available previous information will be the heart of this thesis.

In some cases, it is possible to gain information about the true sales response function even when all that can be observed is the error-loaden terms. Observing a profit of 84 after 12 periods means an average profit (and sales) of 7, which can lead us to believe that a price of 3 results in sales of 7 before the error was introduced. While it is true in this case, it would be seen as ‘too little, too late’, as it took 12 periods to obtain information about a single value. In order to optimize the objective function, one needs to include as much information as possible at every period. That is because the nature of the decision problem at hand allows us to use all the previously gathered data in the decision of the current period. Therefore the early decisions are almost random guesses, while later decisions can be quite well-informed.

Practitioners will be able to use our results in situations with no or little information, such as entering a new market, launching a new product, or indeed, a company has just been founded, like a start-up, and even creating a new market through real innovations. The two main results relevant for application in businesses pertain to advertisement allocation and pricing, both when information is rare.

In Chapter 2, we will analyze how to split a given advertisement budget onto several divisions in order to get the maximum total sales over *all* divisions and the entire planning horizon. Based on theoretical results and the solution in the deterministic case (i.e. without error terms) we construct an algorithm and have it compete against three rules of thumb that are well known and generally used in practice. Correctly interpreting the resulting dataset will require careful analysis.

In Chapter 3, we will determine the best way to set the price of a product in a monopoly to get the maximum profit over the planning horizon. We analyze eight different procedures from the literature and have them compete. We evaluate them based on the market situation.

The dataset that results from the simulation from Chapter 2 has a lot of interesting and unusual properties. This makes answering the question ‘which procedure leads to the best overall sales’ difficult. Therefore, in Chapter 4, we demonstrate the complete derivation of the correct null hypotheses for the dataset.

All proofs to mathematical statements which we derived ourselves will be found in the corresponding chapters or appendices.

## 1.2 Preliminaries, Notation and Definitions

Some of the terms that are used within the next chapters are not explicitly or strictly defined where they appear. We therefore introduce some of them here with a short definition and/or a source. Note also that there is an index at the end of the thesis.

*ANOVA*, see Malhotra (2007), Chapter 16

ANOVA is short for **A**nalysis of **V**ariance. It is a *very* broad topic, so we briefly mention the two applications within this thesis:

1. Given a continuous dependent variable  $y$  (like profits) and only discrete (which in this thesis will always mean ‘taking only finitely many values’, like weekdays) independent variables  $x_1, \dots, x_n$  we can measure the influence of the independent variables  $x_i$  on the dependent variable  $y$ , individually, as well as collectively, as well as the ‘significance’ of that influence. In Chapter 2 we use an ANOVA to find the influence of the response functions and their properties on the objective functions of sales and optimality. In Chapter 3 we find the influence of the response functions on the objective function of forgone profits of the pricing methods.
2. There is a method of performing an ANOVA by means of an OLS estimation. In Chapter 4 we use this method to derive the formula for the null hypothesis of an F-test for an ANOVA with all interaction terms.

*Bhattacharyya coefficient*, see Aherne et al (1997).

Given two probability density functions  $p_1, p_2$  on the same discrete space  $X = (x_1, \dots, x_n)$ , the Bhattacharyya coefficient  $BC$  is defined as

$$BC(p_1, p_2) = \sum_{i=1}^n \sqrt{p_1(x_i) \cdot p_2(x_i)}. \quad (1.6)$$

It is a measure for similarity of the two density functions with  $0 \leq BC \leq 1$ , where  $BC = 1 \iff p_1(x) = p_2(x) \forall x \in X$ , i.e. when the densities are equal.

Cost function

The cost function, also known as the *variable cost of production* is the amount of money a company has to spend on producing a single unit of its goods. Throughout the thesis, cost functions will be assumed to be constant, and known.

Dummy-coding of discrete variables, see Greene (1993), Chapter 8

Given a discrete variable  $X$ , it can only take a finite number  $n$  of possible states  $s_1, \dots, s_n$ . For dummy-coding, we choose one of these states  $s_j$  as a reference, often called the *reference category*. For every  $i \neq j$  of the  $n - 1$  other states we define a dummy variable  $x_i$ . To incorporate the information about the state of a discrete variable into a model, we define them as

$$x_i = \begin{cases} 1 & \text{if } X \text{ is in state } s_i \\ 0 & \text{else.} \end{cases} \quad (1.7)$$

The information that  $X$  is in the reference state  $s_j$  is encoded when every variable  $x_i$  is zero.

Elasticity, see Hruschka (1996) P. 17

Elasticity is a common tool used in economics to break the behavior of a function down to a single number. It is closely related to the derivative, which appears as a factor in it. Given a function  $f$  differentiable in  $x_0$  with  $f(x_0) \neq 0$  we can define the *point elasticity*  $\varepsilon_{f,x_0}$  of  $f$  in  $x_0$  as

$$\varepsilon_{f,x_0} = \frac{\partial f}{\partial x_0} \frac{x_0}{f(x_0)}. \quad (1.8)$$

Often, the analytical definition of a function is unknown, and therefore the actual derivative cannot be determined. In this case it can still be estimated with two pairs of points using the *arc elasticity*: given the points  $(x_0, y_0 = f(x_0))$  and  $(x_1, y_1 = f(x_1))$  with  $y_1 \neq 0, x_0 \neq x_1$  we define the arc elasticity  $\varepsilon_f$  as

$$\varepsilon_f = \frac{\Delta y}{\Delta x} \frac{x_1}{y_1} = \frac{y_1 - y_0}{x_1 - x_0} \frac{x_1}{y_1}. \quad (1.9)$$

F-test, see Malhotra (2007), P. 553-554

We use three kinds of F-tests in this thesis, all of them coming from the following idea: In a linear model, we want to find out if certain linear combinations of coefficients are ‘significantly different’ from being equal to zero.

1. Claiming that every coefficient is already zero will be called a full F-test. It compares the given linear model to a model in which no predictors exist. Full F-tests will be used in Chapter 2.
2. Claiming a certain subset of the coefficients are all equal to zero compares the given linear model to the model of the remaining variables with non-zero coefficients. This will also be used in Chapter 2.
3. Claiming a specific linear combination being equal to zero can be used to compare the effects of certain predictors. It will be used in Chapter 2, and the corresponding linear combinations will be derived in Chapter 4.

Heuristic, see Gigerenzer and Gaissmaier (2011)

Taken from Gigerenzer and Gaissmaier, we can define a heuristic as ‘a strategy that ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods’. Given an optimization problem, such as those presented in Chapter 2 and 3 it is often not possible or not worthwhile to find an exact mathematical solution. A heuristic is a method that gives good results and is constructed to be simple.

Interval

The notation of intervals, i.e. connected subsets of the real numbers will be  $[a; b]$  if the boundary elements  $a$  and  $b$  are included and  $(a; b)$  if they are not, with the obvious extension to the mixed cases. We choose ‘;’ as a separator because ‘,’ is used as a digit grouping symbol, e.g.  $10^3 = 1,000$  and ‘.’ is used as a decimal point, e.g.  $10^{-3} = 0.001$  and we wish to avoid ambiguity.

Lambert-W-function, see Corless et al (1996)

Much like the natural logarithm is used to solve the equation  $y = \exp(x)$ , the Lambert-W-function is used to solve the equation  $y = x \cdot \exp(x)$ . In other words, just like  $\ln(\exp(x)) = x$ , we have  $W(x \cdot \exp(x)) = x$  for certain values of  $x$ . It will be used in Chapter 3 to solve certain functions for the optimal price.

Monte Carlo, see Kalos and Whitlock (2009)

In brief, a Monte Carlo simulation uses samples of a probability distribution to create error terms, and uses the the values of functions loaded by these errors to obtain information about the functions. A common example is an estimation of  $\frac{\pi}{4}$  by sampling points in a square with an inscribed quarter circle.

Natural numbers  $\mathbb{N}$

When we use the natural numbers, we generally mean all non-negative integers, i.e. we include zero. We use  $\mathbb{N}^+$  for the positive integers, excluding zero.

OLS, see Greene (1993), Chapter 6

OLS is short for **O**rdinary **L**east **S**quares, and is an estimator for linear dependencies. Much like ANOVA, it is a *very* broad topic. It tries to find the slope coefficients when assuming a linear relation between one dependent variable  $y$  and one or more independent variables  $x_1, \dots, x_n$ . These models are called *linear models*. In the simplest case it estimates the coefficients  $\beta_0, \beta_1$  of the regression line  $y = \beta_0 + \beta_1 x$  when given a dataset of pairs  $(x^i, y^i)$ .

In Chapter 2 we use increasingly complex linear models by adding the products of the independent variables, which are called *interactions*. When all the dependent variables are discrete, there is an equivalence between OLS and ANOVA, which is used in Chapter 4. Furthermore, one of the methods presented in Chapter 3 uses the OLS to estimate a linear relationship between price and sales.

Robustness, see Huber (1981)

By robustness in this thesis we mean insensitivity to deviations from the assumptions. In the context we use it in Chapter 2, we deviate from the assumption that 20 repetitions are necessary, as results remain valid with 5 repetitions and Bonferroni correction. In Chapter 3 we deviate from the assumption that the tests should be conducted on a 5% level, as the results remain valid when deviating to different levels.

Tukey-test, see Tukey (1949)

The Tukey-test is used to compare the effect of two variables within an ANOVA. Given a significance level  $\alpha$ , it tests whether the means of two cells of the ANOVA differ on that significance level  $\alpha$  by testing if their difference is different from zero. We use Tukey-test in Chapter 3 to compare the pricing procedures pairwise. We can see which procedure is better from the sign of the difference, and we fill certain tables with 'points' whenever one procedure outperforms the other at the specified level  $\alpha$ .

### 1.3 Acknowledgments

This work would not have been possible without the aid and support of the following people:

I would like to thank my primary supervisor, Prof. Dr. Harald Hruschka, for guiding my research and being a constant aid for all of the articles and the thesis.

I would like to thank my secondary supervisor, Prof. Dr. Guido Schryen, for giving additional support, and inspiring the algorithm used for the allocation problem.

I would like to thank my coworkers Dagmar Uhl-Rückerl, Dr. Nadine Schröder and Dr. Andreas Falke for supporting me throughout my time at the chair.

Furthermore, I express my gratitude to the entirety of the University of Regensburg and all its employees.

My deepest gratitude goes to my parents, who encouraged me to take the paths I have chosen and not be deterred.

A special thanks goes out to the student council of chemistry from the University of Regensburg.

Finally, I'd like to thank all the friends who have stayed with me and supported me throughout the years.



## Chapter 2

# Allocation Procedures

In this chapter, we present our solution to an allocation problem with error terms and compare it to rules of thumb frequently used by practitioners. It was accepted for review several journals and is, by the time of submitting this dissertation, under review in ‘Journal of Business Economics’. The paper has two authors: Prof. Dr. Harald Hruschka and me. Changes compared to the original submission consist mainly of layout changes, and deletion of sections necessary only to the journals. The title of the paper as submitted to ‘Journal of Business Economics’ was:

### **Resource Allocation Procedures for Unknown Sales Response Functions with Additive Disturbances**

#### **ABSTRACT**

We develop an exploration-exploitation algorithm which allocates a fixed resource (e.g., a fixed budget) to several units with the objective to attain maximum sales. This algorithm does not require knowledge of the form and the parameters of sales response functions and is able to cope with additive random disturbances. Note that additive random disturbances, as a rule, are a component of sales response functions estimated by econometric methods. We compare the developed algorithm to three rules of thumb which in practice are often used to solve this allocation problem. The comparison is based on a Monte Carlo simulation for twenty replications of 384 experimental constellations, which are obtained from four function types, four procedures (including our algorithm), similar/varied elasticities, similar/varied saturations, high/low budgets, and three disturbance levels. A statistical analysis of the simulation results shows that across a multi-period planning horizon the algorithm performs better than the rules of thumb considered with respect to two sales-related criteria.

## 2.1 Introduction

Allocation decisions in marketing refer to decision variables like advertising budgets, sales budgets, sales force sizes, and sales calls which are allocated to sales units like sales districts, customer groups, individual customers, and prospects. Studies using optimization methods and empirical sales response functions provide evidence to the importance of such allocation decisions. These studies demonstrate that sales or profits can be increased by changing allocation of budgets, sales force or sales calls (Beswick and Cravens (1977); LaForge and Cravens (1985); Sinha and Zoltners (2001)). The average increase of profit contribution across studies analyzed in the review of Sinha and Zoltners (2001) compared to the current policies was 4.5 % of which 71 % are due to different allocations and 29 % are due to size changes. The smaller second percentage can be explained by the well known flat maximum principle (Mantrala et al (1992)).

These studies all require knowledge of the mathematical form of sales response functions which reproduce the dependence of sales on decision variables. In addition, they require that parameter values of sales response functions are available, e.g., determined by econometric methods using historical data or by means of a decision calculus approach which draws upon managers' experiences (Gupta and Steenburgh (2008)). Of course, there are situations in which both econometric methods and decision calculus cannot be applied. Lack of historical data (e.g., for new sales units), lack of variation of past allocations, lack of experiences with the investigated or similar markets constitute possible causes.

In such difficult situations the question arises how management may arrive at rational allocation decisions nonetheless. To our knowledge, the only relevant approach was developed by Albers (1997). For the original German version of this paper, see Albers (1998), which has the same content but was published in a German journal, rather than being a conference paper. Albers demonstrates that, in spite of the lack of knowledge on functional form and parameters, the allocation problem for one resource may be solved by a heuristic which uses elasticity estimates computed from sales and allocation variations of previous periods (i.e., iterations). Albers investigates several sales response functions with different parameter values. Note that these functions and their parameters are not used by the heuristic, they only serve to generate values of the dependent variable sales by deterministic simulation.

To our knowledge, a modified version of this heuristic has been applied in two related publications. The modification deals with a less difficult situation in which elasticities are known and constant across periods. Albers (1997) on the other hand considers the more demanding problem with unknown elasticities which in addition are updated in each period. One of these publications contains an experimental simulation study which adds a growth function in order to reproduce product life cycle effects. This study also considers a second resource to be allocated, but only investigates one function, namely the multiplicative function (Fischer et al (2011)).

The other publication deals with two functional forms, multiplicative and modified exponential (Fischer et al (2013)). Sales generated by simulation in this publication are based on models with true parameter values. These models do not include an overall disturbance term and therefore are still deterministic. Opposed to Albers (1997) it is assumed that elasticities are known, but are subject to error. These erroneous

Method	Advantages	Issues
First rule of thumb (see Section 2.5 for definitions)	stable results	lower on average
Second rule of thumb	may achieve values close to the optimum	leads to unusable results in some constellations, lower on average
Third rule of thumb	works well within a certain budget range	does not scale well with low and high budgets
Iteration along elasticities only as in Albers (1997)	usually converges for deterministic response functions	cannot handle stochastic response functions
Iteration with smoothing and projection (exploration phase) only	useful allocation results for stochastic response functions	inferior to rules of thumb
Developed exploration-exploitation algorithm	performs better than the rules of thumb even for stochastic response functions	possibly first such method, room for improvement

Table 2.1: Comparison of procedures

parameter values are drawn from a probability distribution centered at true values. Allocation problems are solved alternatively by the modified heuristic and numerical optimization. The numerical optimization uses the erroneous parameter values only. By contrast, the modified heuristic also considers generated sales of the previous period (i.e., iteration) in addition. Therefore the modified heuristic performs better than the numerical optimization as the latter lacks any feedback mechanism with respect to previous sales.

Summarizing we note that the simulations presented by Albers (1997), Fischer et al (2011), Fischer et al (2013) generate deterministic sales values, as their sales response models do not include an overall disturbance term. Of course, ignoring disturbances goes against econometrically estimated sales response models which always comprise such a component (for an excellent overview see Hanssens et al (2001b)). That is why we decide to remove the limitation to deterministic sales response models and include additive random disturbances.

We pursue the following research goals in this study. Firstly, we investigate the performance of the iterative heuristic of Albers (1997) to solve the allocation problem, if sales are not deterministic but are generated from sales response functions with additive disturbance terms. Secondly, we develop an alternative iterative algorithm for non-deterministic sales. This algorithm has no knowledge about functional form, parameter values and size of random disturbances. Finally, similar to Albers (1997) we compare to three rules of thumb using several functional forms (i.e., multiplicative, modified exponential, ADBUDG) to generate sales by simulation.

In Section 2.7 we show that the investigated rules of thumb, which are frequently used by practitioners, have several shortcomings which can be overcome by switching to a different procedure. In particular this is a method that practitioners can not only

easily implement, but also modify according to their own particular allocation situation. We also explain how some of the rules of thumb can lead to unusable results, see Table 2.1.

The developed alternative algorithm consists of two stages, exploration and exploitation. March (1991) seems to be the first author combining the concepts of exploration and exploitation. The idea is to divide the planning horizon into two stages. In the first stage data are generated to explore the shape of the response functions. In the second stage the knowledge of their shape is exploited to find optimal values. In each of several iterations we approximate the unknown functions by quadratic polynomials and obtain a solution of the allocation problem by quadratic programming.

To understand the problem setting it may be beneficial to keep the following two examples at the back of one's head while reading: A manufacturer sells and advertises in eight different countries. At the beginning of each period (say month or week), the total advertising budget is allocated onto the eight countries, and at the end of the period, the total sales figures from each country are reported. As another example think about the allocation of a given number of total sales calls (=visits by sales representatives) to several sales districts. Decision makers want to optimize sales/profit in both examples. Customers in each country or sales district might react differently, and few data (if any) are known from previous periods.

The managerial implications are twofold: upon entering a new market with no or very little prior information, the developed algorithm leads to stable results better than those provided by any rule of thumb. Furthermore, in any situation it yields better results independent of the circumstances of the market, *in particular* the size of random disturbances.

Section 2.2 describes the problem from a mathematical point of view. In Section 2.3 we present the necessary preparations for the simulation study, the algorithm is described in Section 2.4. Section 2.5 explains the investigated rules of thumb. The hypotheses shown in Section 2.6 all maintain that the developed algorithm is superior to these rules of thumb. Results of the simulation study are presented and discussed in Section 2.7. In Section 2.8, we investigate the performance of the algorithm under conditions different from those in the simulation study. We also mention several extensions of the allocation problem which may be solved by modifications of our algorithm. The algorithm is presented as pseudocode in 2.9 Appendix A. Further evaluation results not given in the main text can be found in 2.10 Appendix B. Supplementary notes on the generation of sales response functions can be found in 2.11 Appendix C.

## 2.2 Decision Problem

A (scarce) resource  $B$  needs to be allocated to  $n \in \mathbb{N}_{>1}$  units. We have one sales response function  $f_i$  for each unit  $i = 1, \dots, n$  which only depends on its allocated input  $x_i$ . The sales response function maps the allocated budget at the beginning of the period to the obtained sales at the end of the period. Inputs must be non-negative and lower than the resource, i.e.,  $0 \leq x_i \leq B$ .

In addition, the sum of all inputs must not exceed the resource:

$$\sum_{i=1}^n x_i \leq B \quad (2.1)$$

An  $n$ -tuple  $(x_1, \dots, x_n)$  satisfying (2.1) is called an *allocation*. Total sales, i.e., the sum of sales across all units  $\sum_{i=1}^n f_i(x_i)$ , represent the *objective* of this allocation. The goal is to find an allocation maximizing total sales:

$$\max \sum_{i=1}^n f_i(x_i) \quad (2.2)$$

We focus on sales as objective as opposed to profits for several reasons. Firstly, for the sake of comparability with Albers (1997), who also considers sales. Secondly, if the resource is fixed, sales and profits should be highly correlated, as the only difference is the profit contribution for each unit.

Thirdly, it makes calculations and interpretations easier. We doubt that changing the objective to profit will yield more advantages than problems.

As we assume that all sales response functions are monotonically increasing (thereby excluding effects such as supersaturation), we conclude that condition (2.1) is binding and can therefore be rewritten as:

$$\sum_{i=1}^n x_i = B \quad (2.3)$$

**Example 2.2.1.** We now introduce an example which we extend throughout this article whenever a new concept is introduced. Note how an allocation of a budget  $B$  on two units still has one binding, solvable restriction (2.3) and is therefore a one-dimensional optimization problem. Our example therefore has three units, to make sure the problem is not solvable with elementary one-dimensional methods. Furthermore we round the numbers quite abruptly, indicated by ‘ $\approx$ ’, to make the examples more easily readable.

We assume our company runs advertisements in three different countries, and has a total advertising budget of  $B = 6$  that needs to be allocated. If we know nothing about the market, we might choose equal allocations  $(x_1 = 2, x_2 = 2, x_3 = 2)$  which satisfy (2.3).

For each of the three countries we have an advertising response function  $f_i$  that transforms the input  $x_i$  to the sales revenue of the country at the end of that month. In

our example, these functions will be

$$f_1(x) = 5\sqrt[3]{x}; \quad f_2(x) = 3\sqrt[8]{x}; \quad f_3(x) = 3\sqrt[8]{x}.$$

We know these functions are monotonically increasing, and therefore the condition (2.1) is binding and becomes (2.3). The objective value in this case is

$$\sum_{i=1}^n f_i(x_i) = f_1(2) + f_2(2) + f_3(2) = 5\sqrt[3]{2} + 3\sqrt[8]{2} + 3\sqrt[8]{2} \quad (2.4)$$

$$\approx 6.3 + 3.3 + 3.3 = 12.9. \quad (2.5)$$

The decision problem presented so far makes it possible that no (in other words: zero) resources are allocated to a unit (e.g., to make no sales calls in certain districts). Of course, decision makers may find it inappropriate to deprive a unit of all resources. To cope with such a situation, one defines a new problem in which a modified total resource  $B' := B - \sum_{i=1}^n lb_i$  is allocated to the units with functions  $g_i(x_i) := f_i(lb_i + x_i)$  where  $g_i$  is  $f_i$  shifted by the lower bound. Now the inputs of some (or all) units have a lower bound  $lb_i \geq 0$ , i.e., the minimal amount of resource allocated to unit  $i$  and  $B'$  must be zero or positive. One can henceforth allocate onto the functions  $g_i$ , however to simplify the notation we will still refer to the functions as  $f_i$ . In other words, without loss of generality, we may assume the lower bounds of  $f_i$  and hence the lowest possible value of each  $x_i$  to be 0.

**Example 2.2.2.** Assume now, that the budget is actually  $B = 9$ , and we have to allocate at least  $lb_i = 1$  to every unit. All we have to do is allocate the remaining budget

$$B' = B - \sum_{i=1}^3 lb_i = 9 - (1 + 1 + 1) = 6 \quad (2.6)$$

onto the shifted functions  $g_i(x_i) := f_i(lb_i + x_i) = f_i(1 + x_i)$

Albers (1997) derives the following optimality conditions for this decision problem:

$$x_i = \frac{f_i(x_i)\varepsilon_i}{\sum_j f_j(x_j)\varepsilon_j} B \quad \text{for } i = 1, \dots, n \quad (2.7)$$

$\varepsilon_i$  denotes the point elasticity of allocation  $x_i$  which is defined as  $\frac{\partial f_i}{\partial x_i} \cdot \frac{x_i}{f_i}$ . Equation (2.7) is derived from a Lagrangian and is therefore difficult to motivate ad hoc. However it is still easily seen that the  $x_i$  do add up to the total resource  $B$ .

**Example 2.2.3.** The functions in our example are *multiplicative* and hence have constant elasticity, in this case  $\varepsilon_1 = 1/3, \varepsilon_2 = \varepsilon_3 = 1/8$ . Taking the sales values  $f_i$  from the allocation (2,2,2) from Example 2.2.1, the denominator of (2.7) becomes

$$6.3 \cdot 1/3 + 3.3 \cdot 1/8 + 3.3 \cdot 1/8 \approx 2.9,$$

which leads us to the next allocation of

$$x_1 = \frac{6.3 \cdot 1/3}{2.9} B = 0.72 \cdot 6 \approx 4.3, \quad x_2 = x_3 \approx 0.85.$$

Indeed, we see that the new objective value is

$$f_1(4.3) + f_2(0.85) + f_3(0.85) \approx 8.1 + 2.9 + 2.9 = 13.9 > 12.9$$

## 2.3 Simulation

In order to analyze the performance of allocation rules, a Monte Carlo simulation is performed. We choose experimental conditions as close as possible to the original paper presented by Albers (1997).

The successive steps of the simulation can be characterized as follows:

- select one of four functional forms,
- choose one of two budget levels,
- determine for each sales unit parameters, which depend on (non-) similarity of elasticities and saturation levels,
- select one of four procedures which determine the allocations to sales units,
- compute deterministic sales given functional form, parameters and allocations,
- select one of three disturbance levels and add appropriate disturbances to obtain total stochastic sales for each unit.

The simulation comprises 384 constellations, which result from four function types, four allocation procedures, three disturbance levels, and two budget, elasticity and saturation levels. The simulation is replicated twenty times for each constellation. The values for these levels will be given in Section 2.3.2.

Within our simulation, we will make the following assumptions:

- The response functions are monotonically increasing, and mostly concave, the only exception being the S-shaped ADBUDG function, which is convex beyond its inflection point.
- The budget is constant across all periods.
- $\frac{\partial f_i}{\partial x_j} = 0 \quad \forall i \neq j$ . In particular, the response units are mutually independent and the objective function is separable.
- There are no lag-effects, i.e. the response value does not depend on the value of the previous periods
- The error terms added to the response functions are normally distributed with mean 0.

Note that the last assumption is also the usual one made in the literature on nonlinear regression models (see Bates and Watts (1988); Seber and Wild (1989); Cook and Weisberg (1999)).

In the next section we explain which functional forms of sales response functions serve to compute sales, how parameters for sales units are determined based on properties of the functions (with respect to the (non-)similarity of elasticities and saturation levels), and how disturbances are generated.

### 2.3.1 Sales Response Functions

We consider the same four functional forms investigated in Albers (1997), i.e., the multiplicative, the modified exponential, the concave and the S-shaped ADBUDG functions. The deterministic parts of these functions can be written as follows:

$$f_{mul}(x) = ax^b \quad (2.8)$$

$$f_{exp}(x) = M_{exp}(1 - \exp(-xh)) \quad (2.9)$$

$$f_{adc}(x) = M_{adc} \frac{x^{\phi_c}}{G_c + x^{\phi_c}} \quad (2.10)$$

$$f_{adS}(x) = M_{adS} \frac{x^{\phi_S}}{G_S + x^{\phi_S}} \quad (2.11)$$

where  $a, b, M_{exp}, M_{adc}, M_{adS}, h, G_c, G_S, \phi_c, \phi_S$  all are positive parameters.

We now mention properties of these four functional forms (for more details see, e.g., Hanssens et al (2001b)).  $M_{exp}, M_{adc}, M_{adS}$  symbolize maximum values of sales, in other words sales potentials. Given certain parameter restrictions, the first three functions allow for a concave shape, i.e., they reproduce positive marginal effects which are decreasing with higher values of  $x$ . These restrictions are  $0 < b < 1$  for the multiplicative function,  $h > 0$  for the modified exponential function and  $0 < \phi_c < 1$  for the first version of the ADBUDG function. The second version of the ADBUDG function  $f_{adS}$  leads to an S-shape for  $\phi_S > 1$ . S-shaped functions consist of two sections, separated by an inflection point. The first section is characterized by increasing positive marginal effects, the second one by decreasing positive marginal effects.

### 2.3.2 Determining Parameters of Sales Units

For our simulation, we need to assign values to the parameters of the functions in 2.3.1 which we accomplish by demanding certain mathematical properties from our functions. Given a functional form from 2.3.1, we need to construct eight sets of parameters (one for each function) with two such properties: The point elasticity at the optimal allocation should have a predetermined value, and the functions' saturation levels should also take predetermined values. In order to observe possible effects of elasticity level and saturation level, we allow two different sets of values in the following way:

Observe Table 2.2, it is constructed inspired by Table 1 in Albers (1997), which can be found in Appendix 2.11.1 together with a discussion of the differences.

We begin with an experimental condition, say 'low budget, multiplicative form, varied elasticity, similar saturation'. Then we take eight multiplicative functions from 2.3.1 and determine their coefficients such that given a low allocation budget of 1,000,000, the point elasticity of each function at the optimal allocation is one of the eight values from column 3 of Table 2.2 and the saturations of the eight functions are the values in column 4 of Table 2.2. For an actual calculation, see Example 2.3.1 below.

This allows us to add the two factors 'elasticity' and 'saturation' to our experimental design. If they are 'similar', their values are within a close range, so an elasticity around



sales unit	similar elasticities	varied elasticities	similar saturation levels	varied saturation levels
1	0.26	0.11	6,100,000	4,500,000
2	0.27	0.12	6,200,000	10,000,000
3	0.28	0.13	6,300,000	4,500,000
4	0.29	0.14	6,400,000	10,000,000
5	0.31	0.47	6,600,000	4,500,000
6	0.32	0.48	6,700,000	10,000,000
7	0.33	0.49	6,800,000	4,500,000
8	0.34	0.50	6,900,000	10,000,000

Table 2.2: Function Properties for Parameter Generation

0.3 and saturation around 6,500,000 , while in the ‘varied’ case, four functions will have a low level, and four will have a high level.

For the multiplicative function, matters are simple, as  $b$  is the elasticity, and we set  $a := M \cdot B^{-b}$ , where  $M$  is the saturation level.

For the other functions, we perform an iterative search for parameter values. As the saturation level is given, only the remaining parameters need to be determined.

One can derive the elasticity of the modified exponential function as

$$\varepsilon = \frac{\exp(-hx)hx}{1 - \exp(-hx)} \quad (2.12)$$

which can be transformed into a fixed-point formula

$$h = \frac{-\log\left(\frac{\varepsilon(1 - \exp(-hx))}{hx}\right)}{x}. \quad (2.13)$$

Using a starting value  $h_0 > 0$  and  $x = B/8$ , fixed-point iterations quickly converge to a positive value for  $h$  which satisfies the conditions in Table 2.2.

For the ADBUDG-functions which have an additional parameter, we choose special values which yield reasonable shapes and ensure stability of the iteration process. The values are obtained by trial and error and controlled visually (see 2.11.2 Appendix C for a discussion and explanation as to why this is necessary.).

**Example 2.3.1.** In our example so far we used three multiplicative functions, whose coefficients were predetermined. In our simulation study, we want the functions *properties* to be fixed, in order to *derive* the coefficients.

Starting from the desired properties of  $\varepsilon = 1/3$  and  $f(6) \approx 9$ , we set the exponent  $b$  to be the elasticity  $b = \varepsilon = 1/3$  and  $a = M \cdot B^{-b} = 9 \cdot 6^{-1/3} \approx 5$  and we obtain the first function of Example 2.2.1.

### 2.3.3 Disturbances

We add normally distributed disturbances  $u \sim \mathcal{N}(0, \sigma^2)$  to the deterministic part of each function. Variances  $\sigma^2$  are set to attain a desired share of explained variance (i.e.,

$R^2$  value) for the dependent variable sales. We consider additive normally distributed disturbances as these are included in the majority of nonlinear regression models. Switching to, say, multiplicative error terms might give the algorithm we propose an unfair advantage as it is based on elasticities. The desired  $R^2$  values amount to 0.9, 0.7, and 0.5. For each function, the error variance  $\sigma^2$  is set to the value which leads the  $R^2$ -value closest to its desired value in a regression of that function using 2,000 values from a discrete uniform distribution of integers as inputs. We choose these values for  $R^2$  as they are easily understood and well known by both practitioners and researchers. To avoid negative outputs ( $f(x) + u$  can be negative when  $f(x)$  is too small), the output was defined as  $\max(f(x) + u, 0)$ . This is a mixed distribution where the discrete value 0 can have positive probability. This only happens in extreme cases, therefore the distribution of sales is very close to a normal distribution.

**Example 2.3.2.** Extending the situation from Example 2.2.3 we add the following standard normally distributed random numbers ( $u_1 = -0.6, u_2 = 0.2, u_3 = -0.8$ ) to the sales function, to obtain

$$f_1(4.3) + u_1 + f_2(0.85) + u_2 + f_3(0.85) + u_3 = 8.1 - 0.6 + 2.9 + 0.2 + 2.9 - 0.8 = 12.7$$

In particular, we see the effect the disturbance has on the objective value, as it is now lower than its value for equal allocations, even though allocation  $(4.3, 0.85, 0.85)$  is better than  $(2, 2, 2)$  in the deterministic case.

Let us illustrate the process of obtaining the correct variances. We now sample 2,000 points between 0 and 6 on the function  $f_1(x) = 5\sqrt[3]{x} + u$ . We take the data of these 2,000 pairs  $(x, f(x))$  and perform a non-linear least squares regression  $f(x) \sim ax^b$ . This regression gives a value for the coefficient of determination  $R^2 \approx 0.67$ . If we want to find the variance such that  $R^2 = 0.5$  we need to increase the variance of the distribution we are sampling from. If we want to find the variance such that  $R^2 = 0.7$  we need to decrease it.

## 2.4 Developed Algorithm

Albers (1997) intends to show that in allocation problems an iterative heuristic along elasticities always outperforms rules of thumb, independent from functional form, other properties of the functions and correlations of starting values with optimal allocations. He considers for each of the four functions discussed in the previous section eight constellations of different values of properties and correlations. Sales are computed on the basis of these functions without adding disturbances though the latter are a component of most econometric models. Results obtained by his iterative algorithm are compared to those computed by the three rules of thumb which we explain in Section 2.5 below. Overall, his iterative algorithm outperforms all rules of thumb by far.

In our study the iterative algorithm of Albers turns out not to work well for non-deterministic sales response functions with additive disturbances. Disturbances which directly affect sales cause elasticities and new allocation values to fluctuate. These fluctuations can, among other problems, lead to negative slopes and hence negative elasticities.

To overcome these problems we use a modified exploration-exploitation algorithm. We call it modified, as the exploration phase does not aim to explore the entire response surface itself. It only looks at a concave section which is then approximated by a parabola.

### 2.4.1 Exploration

For exploration we modify the iterative algorithm of Albers (1997). Estimated elasticities are heavily influenced by the additive disturbances. As the analytical solution requires elasticities near the optimum, it is worthwhile to emphasize more recent data points. We therefore perform first order exponential smoothing in the following way:

$$\tilde{\epsilon}_t := (1 - \beta)\tilde{\epsilon}_{t-1} + \beta\hat{\epsilon}_t \quad (2.14)$$

with  $\beta \in [0, 1]$ . We use  $\beta := 0.85$  based on a Monte Carlo simulation with  $\beta$  as a regressor with discrete values between 0.05 and 1. However, other values can be used, based on how much influence of the previous period is preferred.  $\tilde{\epsilon}_t$  is the smoothed elasticity and  $\hat{\epsilon}_t$  the estimated elasticity in period  $t$ .

The other modification consists in projecting each elasticity value into the interval  $[0.01, 0.5]$ , i.e., if the calculated value is above 0.5 it is set to 0.5, and to 0.01 if it is below 0.01. This value range conforms with the results of meta-analyses of advertising elasticities (Assmus et al (1984), Sethuraman et al (2011)). Again, in an actual application a practitioner may choose any interval reasonable based on his expectations on the limits of the elasticity.

In its exploration stage the algorithm generates data points close to the optimum for a fixed number of periods. But elasticities still jump around too much. Therefore, there is dire need for a method which dampens disturbances.

**Example 2.4.1.** Returning to the initial Example 2.2.1, we might deduce that unit 1 is more profitable than the other two, and suggest a second allocation of  $(4, 1, 1)$ , which

then (in the deterministic case) yields an objective value of

$$f_1(4) + f_2(1) + f_3(1) = 5\sqrt[3]{4} + 3\sqrt[8]{1} + 3\sqrt[8]{1} \approx 8 + 3 + 3 = 14.$$

As the functional form is invisible to the allocation methods, we now need to estimate the arc elasticities, using the formula

$$\hat{\epsilon}_i := \frac{\Delta f_i}{\Delta x_i} \frac{x'_i}{f'_i}$$

where  $f'_i$  and  $x'_i$  denote the values in the second period of unit  $i$ . We obtain

$$\hat{\epsilon}_1 = \frac{7.94 - 6.3}{4 - 2} \frac{4}{7.94} \approx 0.41, \hat{\epsilon}_2 = \hat{\epsilon}_3 \approx 0.09$$

which is close enough to the actual elasticities of  $\epsilon_1 = 1/3, \epsilon_2 = \epsilon_3 = 0.125$ . For the next allocation we use conditions (2.7) to obtain

$$x_1 = B \frac{\hat{\epsilon}_1 f'_1}{\sum_{i=1}^3 \hat{\epsilon}_i f'_i} = 6 \cdot \frac{3.27}{3.82} = 5.15, x_2 = x_3 = 0.43,$$

which in turn yields a new objective value of

$$f_1(5.15) + f_2(0.43) + f_3(0.43) \approx 8.63 + 2.7 + 2.7 = 14.03$$

and we estimate the next elasticities. For example we obtain

$$\epsilon_1 = \frac{8.63 - 7.94}{5.15 - 4} \frac{5.15}{8.63} = 0.36.$$

Had that value been higher than 0.5, we would have set it to 0.5. Before calculating the next allocation, we perform exponential smoothing by setting

$$\tilde{\epsilon}_1 = (1 - 0.85) \cdot 0.41 + 0.85 \cdot 0.36 = 0.3675.$$

### 2.4.2 Exploitation

The general idea of the exploitation stage of the algorithm can be easily understood when looking at modified exponential functions with varying parameters. Figure 2.1 shows an example of the modified exponential function for one of the eight sales units upon which the budget of 8,000,000 is to be allocated, where  $x$  is the budget allocated to that unit and  $y$  its sales value. The optimal allocation to the unit in the example drawn in Figure 2.1 is around 640,000. Nevertheless, even when disturbances are small, the algorithm and each of the three rules of thumb still fluctuate a lot, showing no sign of stability, although they do not leave a certain interval of the domain in each variable (and hence of the codomain).

This area, highlighted by a rectangle in the figure, looks like it can be easily approximated by a parabola, i.e., a polynomial of degree two. Assuming the functions were *actually* polynomials of degree two, quadratic programming gives an exact solution,

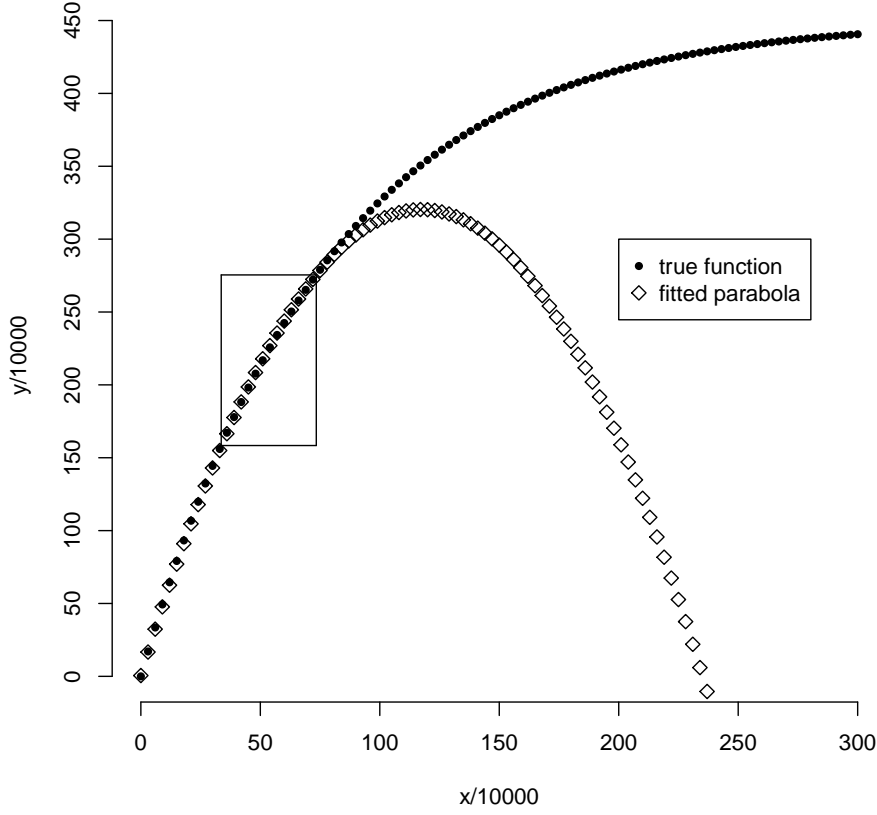


Figure 2.1: True Function and Fitted Parabola

because the total resource restriction (2.3) is linear.

Exploitation consists of two steps. In the first step, for each unit  $i$  a quadratic regression of the form

$$y_i = \gamma_{i,0} + \gamma_{i,1} \cdot x_i + \gamma_{i,2} \cdot (x_i)^2 \quad (2.15)$$

is performed, where  $y_i$  and  $x_i$  are the sales and allocation values for unit  $i$  and  $\gamma_{i,j}$  are the coefficients to be estimated. For each regression, the data consist of all periods observed so far (using only a smaller number of the most recent values does not improve results).

In the second step an allocation which is optimal for these approximations obeying

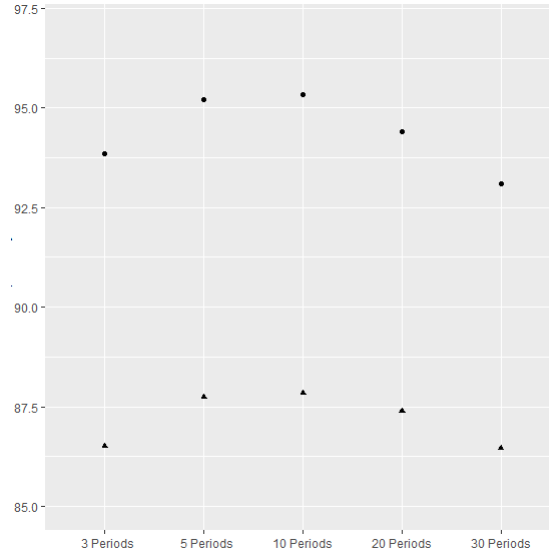
the total resource restriction is determined by quadratic programming, using the method of Goldfarb and Idnani (1983). This allocation and its corresponding sales value constitute an additional data point for the next quadratic regression. This two-step process is repeated until a total of 40 periods is reached, where 40 is the planning horizon within our study.

A problem arises when  $\gamma_{i,2}$  is estimated as a positive number for any sales unit, as the matrix in the quadratic program is no longer positive definite. In particular, this means that the sampled area suggests a convex shape of the sales function, which is problematic. If the function can still be estimated as monotonically increasing however, this can be remedied by setting  $\gamma_{i,1}$  to the slope of the regression line, and setting  $\gamma_{i,2}$  to a negative value close to zero (we choose  $-10^{-15}$ ), thereby forcing the quadratic program to accept something that is basically a straight line rather than a parabola. In the worst-case-scenario, additionally, the slope of the regression line may be negative. This case is very rare and the allocation to this unit will almost certainly be zero. One should remember however, what this actually means: The shape of the data points resembles a monotonically decreasing, convex(!) function and would hence arise either from a few very unfavorable disturbances in a row or an outer influence that cannot be explained by additive disturbances. Surely, in this case, the function should be thoroughly analyzed instead of continuing the application of *any* algorithm or heuristic. The situation will be briefly mentioned in Section 2.8.

We abstain from offering a numerical example in the sake of length and readability, as such an example would not lead to further understanding of the issues.

It is furthermore reasonable to ask whether exploration and exploitation are both necessary, and what the optimal period is to switch from exploration to exploitation.

Figure 2.2 shows the average means of total sales and optimality (defined as ratio of total sales and optimal total sales) for all constellations of the exploration-exploitation procedures, for switches from exploration to exploitation at  $t = 3, 5, 10, 20, 30$  periods. Switching after three periods represents the case of *only exploitation* and switching after 30 periods is essentially *only exploration*. Figure 2.2 implies visually, that neither of these extremes are useful, and a balance between them must be found. For this simulation, we therefore switch from exploration to exploitation at period 10.

Figure 2.2: Means when switching after  $t$  periods

## 2.5 Rules of Thumb

We start from the same rules of thumb as Albers (1997):

1. Allocation proportional to sales of a unit in the previous period.
2. Allocation proportional to sales of a unit divided by its allocation in the previous period.
3. Allocation proportional to the saturation level of a unit.

The third rule cannot be implemented, as saturation levels, like all parameters of functions, are unknown. That is why we replace it by the following rule.

- 3.' Allocation proportional to maximum sales of a unit observed so far.

This is an appropriate substitute for the third rule using maximum sales as proxy of the saturation level.

For the remainder of the paper, we will refer to rules 1, 2 and 3' as first, second, and third rule of thumb, respectively.

**Example 2.5.1.** Starting from the allocation  $(4, 1, 1)$  from Example 2.4.1, the first rule suggests to use as next allocation  $x'_i = B \cdot f_i / (\sum_{i=1}^3 f_i)$  leading to

$$x'_1 = 7.9 / (7.9 + 3 + 3) = 3.42, \quad x'_2 = x'_3 = 3 / 13.9 = 1.29$$

and an objective value of

$$7.53 + 3.1 + 3.1 = 13.73$$

Rule 2 suggests using  $x'_i = B \cdot \frac{f_i}{x_i} / (\sum_{i=1}^3 \frac{f_i}{x_i})$  leading to

$$x'_1 = 1.5, \quad x'_2 = x'_3 = 2.25$$

and an objective value of

$$5.7 + 3.3 + 3.3 = 12.3.$$

The third rule uses the maximal observed values of each unit. For the first unit, that is 7.94, for the second and third unit that is 3.27. So we allocate

$$x_1 = B \cdot \frac{7.94}{7.94 + 3.27 + 3.27} = 3.29, \quad x_2 = x_3 = B \cdot \frac{3.27}{7.94 + 3.27 + 3.27} = 1.35$$

and obtain an objective value of

$$7.44 + 3.11 + 3.11 = 13.66.$$

## 2.6 Hypotheses

To properly assess the data from the simulation, we formulate hypotheses which we test statistically. As we want to determine whether the developed algorithm performs better than the rules of thumb, we construct one hypothesis for each combination of rule of thumb and dependent variable:

**$H_{1,a}$ :** The developed algorithm leads to higher sales than the first rule of thumb

**$H_{1,b}$ :** The developed algorithm leads to higher sales than the second rule of thumb

**$H_{1,c}$ :** The developed algorithm leads to higher sales than the third rule of thumb

**$H_{2,a}$ :** The developed algorithm leads to sales closer to the optimum than the first rule of thumb

**$H_{2,b}$ :** The developed algorithm leads to sales closer to the optimum than the second rule of thumb

**$H_{2,c}$ :** The developed algorithm leads to sales closer to the optimum than the third rule of thumb



## 2.7 Evaluation of Procedures

For the sake of comparing the developed algorithm we conduct a simulation study with 384 constellations, which result from four function types, four procedures, similar/varied elasticities, similar/varied saturations, two budget-levels, three disturbance levels and generate twenty replications for each constellation.

We therefore define the following dummy-coded variables for our models:

“Proc” corresponds to the three rules of thumbs and the developed algorithm, “Form” corresponds to the four types of functions, “Dist” corresponds to the  $R^2$  values of 0.9, 0.7 and 0.5, “Elas” and “Satu” correspond to “similar” and “varied” elasticities and saturations, respectively. “Budg” corresponds to the two budget levels of 1,000,000 and 8,000,000.

The objective function of the allocation problem in each constellation can be written as

$$\begin{aligned} \max \quad & \sum_{t=1}^{40} \sum_{i=1}^8 f_i(x_{t,i}) \\ \text{s.t.} \quad & x_{t,1} + \dots + x_{t,8} = B \quad \forall t \in \{1, \dots, 40\}, \end{aligned} \quad (2.16)$$

Allocations  $x_{t,i}$  may depend on previous sales and allocations, i.e., all values  $f_i(x_{\tau,i})$  and  $x_{\tau,i}$  with  $\tau < t$  may be used to determine the  $x_{t,i}$ .

The performance of procedures is measured by two different dependent variables, “Sales” and “Optimality” which both are computed as arithmetic means across 40 periods. These two dependent variables normalize total sales attained by rules of thumb or the algorithm. We want normalizations to differ with respect to the (non-) consideration of additive random disturbances. That is why the denominator in the definition of “Sales” includes random disturbances, which on the other hand are excluded by the denominator in the definition of “Optimality”.

The first dependent variable “Sales” equals total sales divided by the maximum attained value given a fixed functional form and a fixed budget level over all other constellations. The other constellations result from four allocation procedures, three disturbance levels, and two budget, elasticity and saturation levels. The division to compute “Sales” is necessary as the four function types yield quite different values for total sales (which is especially pronounced in the case of ADBUDG-functions) and the attainable maximum sales depend on the budget level.

The second dependent variable “Optimality” is defined as ratio of total sales and optimal total sales. Optimal total sales are determined by optimizing on the basis of the true response functions without disturbances, i.e., assuming knowledge of sales response functions and their parameters. Optimality therefore shows to what extent a procedure which lacks knowledge of the response functions attains optimal total sales on average. A value of 1.0 for optimality indicates that average total sales as a rule equal their optimal value.

As mentioned above, the S-shaped ADBUDG-function is not concave, and hence a problem arises with local and global optima. In particular, the nine control algorithms that search for the optimal solution may get stuck in local optima. Therefore, optimal-

Dependent Variable	Algorithm	Rules of Thumb		
		1	2	3
Sales	0.8711	0.8522	0.7723	0.86
Optimality	0.9536	0.9305	0.8268	0.9419

Table 2.3: Arithmetic Means of Dependent Variables for Each Procedure

ities greater than 1.0 may be obtained. To avoid such values we additionally divide optimalities by their highest value.

For each of the four procedures Table 2.3 gives arithmetic means of both dependent variables. The algorithm attains the highest (i.e., best) values for both dependent variables, followed by the third rules and the first rule of thumb. The second rule of thumb attains the worst values.

To thoroughly compare the procedures we start from the following two linear regression models comprising main effects only:

$$Sales = \beta_{1,0} + \beta_{1,1}Proc + \beta_{1,2}Form + \beta_{1,3}Dist + \beta_{1,4}Elas + \beta_{1,5}Satu + \beta_{1,6}Budg + e \quad (2.17)$$

$$Optimality = \beta_{2,0} + \beta_{2,1}Proc + \beta_{2,2}Form + \beta_{2,3}Dist + \beta_{2,4}Elas + \beta_{2,5}Satu + \beta_{2,6}Budg + e \quad (2.18)$$

In these equations for  $i = 1, 2$ ,  $\beta_{i,0}$  is the intercept,  $\beta_{i,1}, \beta_{i,2} \in \mathbb{R}^3$ ,  $\beta_{i,3} \in \mathbb{R}^2$ , and  $\beta_{i,4}, \beta_{i,5}, \beta_{i,6} \in \mathbb{R}$  are coefficient vectors, each multiplied by dummy-coded variables, and  $e$  is the usual normally distributed error term.

Reference categories of these dummy-variables are the developed algorithm, the multiplicative function, low disturbances (i.e., a sales response functions with high  $R^2$  values of about 0.9), the low budget, and similar values for elasticities and saturations, respectively.

The estimation results of the main effect models both for “Sales” and “Optimality” are shown in Table 2.4. Coefficients for procedures are in line with the arithmetic means of Table 2.3. They reflect that all rules of thumb perform worse than the developed algorithm (the reference category). These coefficients also indicate that allocation proportional to maximum sales of a unit observed so far (Proc4) performs better than the other two rules of thumb. The second rule of thumb (Proc3), allocation proportional to sales of a unit divided by its allocation in the previous period, clearly turns out as overall worst procedure.

The other coefficients of Table 2.4 indicate whether we obtain higher or lower values of the two dependent variables for a certain category of the respective independent variable. To our opinion most of these results are fairly intuitive. The higher complexity of S-shaped ADBUDG functions (=Form4) leads to lower values than the other functional forms. Both the modified exponential (=Form2) and the concave ADBUDG function are better than the multiplicative function (the reference category). The coefficients for Dist2 and Dist3 reflect that a higher level of disturbances makes it difficult

	Dependent Variable: Sales		Dependent Variable: Optimality	
Variable	Coefficient	t-Value	Coefficient	t-Value
Intercept	0.81	204.130***	1.014	274.947***
Proc2	-0.019	-5.833**	-0.023	-7.687***
Proc3	-0.099	-30.485***	-0.127	-42.108***
Proc4	-0.011	-3.42***	-0.012	-3.889**
Form2	0.042	12.905***	0.023	7.476***
Form3	0.078	23.936***	0.013	4.406***
Form4	-0.135	-41.729***	-0.096	-31.827***
Dist2	-0.011	-3.795***	-0.021	-8.118***
Dist3	-0.038	-13.605***	-0.067	-25.674***
Elas2	0.004	1.884	-0.046	-21.608***
Satu2	0.067	29.382***	-0.036	-17.067***
Budg2	0.091	39.491***	0.05	23.640***
$df = 7668$	$R^2 = 0.5325, R^2_{adj} = 0.5318$		$R^2 = 0.4485, R^2_{adj} = 0.4478$	
* p-value<0.05, ** p-value<0.01, *** p-value<0.001				

Table 2.4: Main Effect Regression Models

to attain good values of both dependent variables. Varied elasticities are accompanied by lower optimality values, but do not have a significant effect on sales. If saturations are varied, sales are higher, but optimality is lower. Higher sales are simply a consequence of upward scaled functions. On the other hand, varied saturations lead to higher derivatives and fluctuations which both make optimization more difficult. Positive and significant coefficients for Budget indicate that higher budgets favour higher values of both dependent variables.

### 2.7.1 Results for Regression Models with Interactions

We also estimate regression models which in addition include certain sets of interaction terms. The most simple of these models includes only pairwise interactions (see 2.10.3 Appendix B).

To consider additional interaction terms, we also investigate models (one for each dependent variable) with all triple, quadruple, quintuple and sextuple interactions. Quite interestingly, the full models with sextuple interactions not only lead to the best variance explanations (adjusted  $R^2$  values amount to 0.9257 and 0.8705 for “Sales” and “Optimality”, respectively), but F-tests also confirm that the full models outperform all the restricted models. A comprehensive discussion can be found in 2.10.1 Appendix B. Because of the superiority of the full models we need a test of differences between the algorithm and each rule of thumb which goes beyond the main effects model by also considering all interactions. We therefore use the method of Gahler and Hruschka (2018b) to construct a vector (=matrix with one row) of multiplicities (= appearances) for the coefficients in the OLS-estimator. The corresponding matrix (whose product with the coefficient vector is zero under the null hypothesis) was the basis of an F-test. Hence a positive product indicates that the algorithm performs better than a rule

of thumb, and the F-statistic reveals the significance of better performance. A more detailed explanation can be found in 2.10.2 Appendix B

For the dependent variable “Sales” we obtain average differences between the algorithm and each of the three rules of thumb of 1.6093, 9.3232, and 0.8757, respectively (the corresponding F-statistics are 176.4363, 5,922.021 and 52.2406). All of these are significant at a level (far) below 0.001, supporting  $H_{1,a}$ ,  $H_{1,b}$  and  $H_{1,c}$ .

For the other dependent variable “Optimality” we obtain differences of 1.9223, 11.8959 and 0.8215, respectively (F-statistics are 198.7278, 7,610.8667 and 36.2925). All of these are significant at a level (far) below 0.001, supporting  $H_{2,a}$ ,  $H_{2,b}$  and  $H_{2,c}$ .

In the analysis above, we use 20 replications as we have a large model and wish to have 20 times more data points than variables.

To answer a question from one of the anonymous reviewers we reduce the number of replications to five. We also perform a Bonferroni correction to take into account that we make multiple comparisons between procedures. The robustness of our results is confirmed by this additional analysis which shows that the effect differences between the rules of thumb and our procedure as described above keep their correct signs (i.e., the effect of the procedure is higher than the effect of each rule of thumb), the corresponding F-tests remain significant below  $\alpha = 0.001$  and the combined significances from the Bonferroni correction remain below 0.001 as well. Therefore, even under these restrictive conditions, the null hypotheses can be rejected, individually and collectively.

Furthermore, issues might be raised concerning the similarity of the results for the suggested procedure and the third rule of thumb. Ignoring for a while that the goal of hypothesis testing is finding whether differences are significant despite magnitude of means, we can take a look at the distribution of the obtained values for optimality and sales after 40 periods as a box-plot. Observe Figure 2.3 containing these box-plots. We see that the third rule of thumb not only has more outliers, but also more extreme outliers. With respect to optimality, there are individual outliers below 50%, i.e. the rule leads to values over the entire planning horizon that are below half of the optimal possible values and quite a lot of the outliers are below 70%. For sales, this holds true for all methods, however the third rule of thumb now has many outliers below 40%.

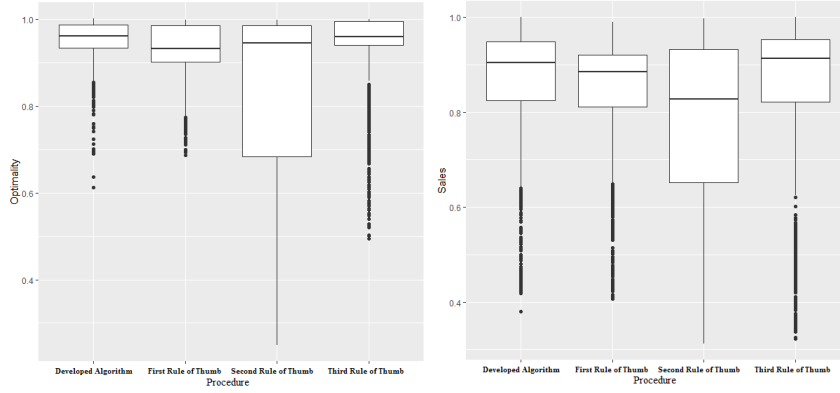


Figure 2.3: Box-Plot of Optimality and Sales

## 2.8 Conclusion

Our study shows that the iterative algorithm of Albers (1997) is not appropriate if sales response functions are not deterministic and include additive disturbances. Disturbances which directly affect sales cause elasticities and new allocation values to fluctuate and sometimes even produce negative elasticities. Let us remind you that elasticities should be positive and lie in a certain interval (see Section 2.4.1). Moreover, its performance does not really improve if elasticity estimates are only projected into this interval. Our iterative algorithm does not suffer from such problems.

A conventional approach begins by estimating a sales response function for each unit based on a given data set of sales and allocations. Then the proper allocation problem is solved by numerical optimization using the estimated sales response functions. Because of its iterative nature our algorithm is different. In the exploitation stage we start with an approximation to the unknown sales response functions for each unit using a data set generated in the exploration stage. These approximations are used as inputs of a quadratic programming problem whose solution provides new allocations to the units. These new allocations and sales due to these allocations are added to the data. Based on the data set enlarged this way, approximations of sales response functions are updated, for which new allocations are determined by quadratic programming and so on for several iterations (i.e., periods).

In each iteration the data are extended by allocations determined by an approximate optimal program and their corresponding sales. Therefore the fit of the approximate response functions in a neighborhood close to the optimal solution gets more important. The conventional approach on the other hand tries to estimate response functions that also fit well for allocations far from the optimal solution.

In addition to the simulation study presented, we also have examined whether performance of the developed algorithm remains stable if it has to deal with different conditions. First of all we have analyzed how procedures behave given a different number of units. The performance of the algorithm remains similar to that obtained in the simulation no matter whether we consider less (four) or more (twenty) units.

Remember that in the simulation study in each experimental constellation sales are generated by functions which have the same form for all units. Of course, in real life situations, sales response of units may be diverse due to, e.g., economic or cultural factors which may differ sharply between regions or customer groups. Now, sales response can no longer be reproduced by different coefficient values and requires the use of different functional forms. Combinatorics tells us right away that an exhaustive proof of the superiority of our algorithm in this case is not possible (as we would need to consider  $4^8 = 65536$  different function constellations), so we have looked at several randomly as well as explicitly designed conditions. We randomly choose the function type for each unit and repeated this exercise by varying the number of units. Still, for these constellations, our algorithm usually determines better solutions than the rules of thumb.

Summing up, we suggest that the developed algorithm should be preferred to the investigated rules of thumb and the iterative algorithm of Albers (1997) if marketing allocation problems of the form shown in Section 2.2 must be solved and the underlying sales response functions are unknown and not deterministic. We justify this suggestion by the good performance of the algorithm demonstrated by the simulation study and its stability under changing conditions.

Future work might consider modifications of the algorithm for the decision problem investigated here. What effect would it have to choose a different algorithm in the exploration stage? While different smoothing variations have been attempted for the exploration phase, none led to improvements. A next goal would be to examine if other algorithms are more suitable for the exploitation stage. This might mean small adjustments of parameters, or a completely new algorithm.

As mentioned in Section 2.4, a problem arises if the parabola has a positive leading coefficient and the regression line has a negative slope. This effect usually occurs if a unit repeatedly receives allocations near zero. Due to the additive disturbances, the data set for the regressions then consists of very similar  $x$ -values, while the  $y$ -values vary a lot. As low allocations are usually not optimal, an amendment of the algorithm may be beneficial for such situations.

Modifying the algorithm to solve more general marketing decision problems also seems to be an interesting task of future research. One extended decision problem results if one or several sales functions may change suddenly, for example due to quality problems, successful advertising campaigns, entering another stage of the product life cycle or the addition or eliminations of allocation units. We suspect that under such circumstances the exploration stage will have to start once again. For situations with gradual change on the other hand an appropriate adjustment would be to delete older data points before each iteration. One could also investigate multi-variable generalizations. In one generalization, allocations affect sales of the same unit as well as sales of other units. Another more challenging generalization allows for marketing variables of different types. Examples of such variables are advertising and price or advertising and sales effort, where both variables have an effect on sales of different units.

## 2.9 Appendix A: The Algorithm in Pseudocode

Notation: for two vectors  $a$  and  $b$  of the same dimension, we denote by  $\langle a, b \rangle$  their dot product, and by  $a * b$  the componentwise multiplication, i.e., the vector  $(a_i b_i)_{i=1}^n$

**Data:**  $x^1 = (x_1^1, \dots, x_n^1)$ ,  $x^2$  start values,  $0 \leq \beta \leq 1$  smoothing parameter, maxit: planning horizon,  
 B: resource,  $I_n$ : identity matrix of dimension  $n$ , exex=10: number of iterations in the  
 exploration stage,  $\tilde{f}$ : multidimensional map consisting of the separate response functions in  
 each component,  $DF$ : array which will be filled with the data points of all functions

```

 $y^1 := \tilde{f}(x^1); y^2 := \tilde{f}(x^2);$ 
while  $i < \text{exex}$  do
  Test( $x^1, x^2, y^1, y^2$ );
   $\varepsilon := \frac{y^2 - y^1}{\frac{x^2 - x^1}{x^1}}$ ; Epstest( $\varepsilon$ );
  if  $i \geq 2$  then
     $\varepsilon_{sm} := \varepsilon \cdot (1 - \beta) + \varepsilon_{old} \cdot \beta$ 
  else
     $\varepsilon_{sm} := \varepsilon$ ;
  end
   $up := \frac{y^1 * \varepsilon_{sm}}{\langle y^1, \varepsilon_{sm} \rangle} \cdot B$ ;
   $\varepsilon_{old} := \varepsilon_{sm}$ ;
   $x^2 := x^1; y^2 := y^1$ ;
  Test2( $up, x^2$ );
   $x^1 := up; y^1 := \tilde{f}(x^1)$ ;
  add ( $x^1, y^1$ ) to  $DF$ ;
end
for  $i = \text{exex}$  to  $\text{maxit}$  do
  for  $j = 1$  to  $n$  do
    Perform Regression  $y_j = \gamma_{j,0} + \gamma_{j,1}x_j + \gamma_{j,2}(x_j)^2$ 
  end
   $D$ : Diagonal Matrix containing values  $-2\gamma_{j,2}$ 
   $A$ :  $n$  by  $n+1$ -Matrix with  $-1$  in the first column followed by  $I_n$ 
   $d$ : Vector containing values  $\gamma_{j,1}$ 
   $bv$ : Vector of length  $n+1$  containing  $(-B, 0, 0, \dots, 0)$ 
   $up := \text{solve.QP}(D, d, A, bv)$ 
   $x^1 := up; y^1 := \tilde{f}(x^1)$ ;
  add ( $x^1, y^1$ ) to  $DF$ ;
end

```

### Algorithm 1: Developed Algorithm

Remark: The functions Test and Test2 check if the entries are too small or too close to each other, Epstest projects elasticities into the Interval  $[0.01, 0.5]$ .

The function solve.QP solves the quadratic program as described in Goldfarb and Idnani (1983) with the notation from the R-Package “quadprog” by Berwin and Turlach (2013).

## 2.10 Appendix B: Additional Regression Results

### 2.10.1 Comparison of Nested Models

While Table 2.4 has entries from which certain statements can be deduced, one should take these values with a grain of salt. The  $R^2$  and adjusted  $R^2$  of these main effect models amount to 0.5325 and 0.5318 for sales and 0.4485 and 0.4478 for optimalities, respectively.

For  $1 \leq k \leq 6$  we therefore define the  $k$ -fold interaction model (or short,  $k$ -fold model) as the linear multiple regression model that contains all interactions up to degree  $k$ , whereby the 1-fold interaction model is just the main effect model and the 6-fold interaction model is the full model. The  $R^2$  and adjusted  $R^2$  of these full models amount to 0.9294 and 0.9257 for sales and 0.877 and 0.8705 for optimalities, respectively.

To be absolutely sure that the full model is better for estimating the effects, we perform statistical tests. For  $1 \leq k < k' \leq 6$ , the  $k$ -fold model is nested within the  $k'$ -fold model. We therefore perform F-tests between the 6-fold model and each  $k$ -fold model nested within, for both dependent variables. For both dependent variables, the 6-fold model is superior to the 5-fold model at a significance level of 0.05. In both cases for all  $1 \leq k < 5$ , the 6-fold model is superior to the  $k$ -fold model at a significance level (far) below 0.001. Note that these tests are in fact necessary, since the F-tests of several nested models do not always behave in a transitive manner (see Cameron and Trivedi (2005) Section 7.2.7).

### 2.10.2 F-Tests in the Full Models

The tests on the full model are conducted by method explained in Gahler and Hruschka (2018b). Following their notation, the primary variable is the procedure variable, and the other predictors are secondary variables.

The conditions necessary for using the method are as follows:

1. The multiple regression model contains only categorical predictors.
2. The regression includes all interaction terms of every possible degree.
3. There are no empty cells, i.e. the constructed design matrix is a generic design matrix in the sense of Definitions 1 and 4 in Gahler and Hruschka (2018b).

The first two conditions are satisfied by construction, the third is satisfied since our design is balanced.

Furthermore it only makes sense to use the method if the explanatory power of the full interaction model is higher than for all nested models. This condition is satisfied as well, as detailed in 2.10.1 Appendix B.

We could use Theorems 2 and 3 from Gahler and Hruschka (2018b) to determine the estimated cell means from the coefficients. Thankfully we can skip this rather tedious procedure by directly applying the result from Section 2.3 from Gahler and Hruschka (2018b), which states that the number of appearances (= the multiplicity) of a coefficient in the null hypothesis is zero if it does not describe the non-reference



category one wishes to compare to, and equal to the product of the number of categories which according to the coefficient are in the reference category otherwise.

Therefore, when comparing the algorithm to the first rule of thumb, coefficients which represent an interaction term wherein the category of first rule of thumb does not appear (for example the coefficient of  $\text{Proc}_3 \cdot \text{Elas}_2$ ), have a multiplicity of zero. The multiplicity of all the other coefficients equals the product of the number of categories for those variables which assume a reference value. For example, the cell where  $\text{Proc}=\text{ROT1}$ ,  $\text{Form}=\text{multiplicative}$ ,  $\text{Dist}=0.9$ ,  $\text{Elas}=\text{similar}$ ,  $\text{Satu}=\text{varied}$  and  $\text{Budg}=\text{low}$  has four variables (Form, Dist, Satu and Budg) in their reference categories. The multiplicity is therefore the product of their numbers of categories:  $4 \cdot 3 \cdot 2 \cdot 2 = 48$ .

This way the multiplicity of each coefficient can be determined. The product of this vector of multiplicities and the OLS estimated coefficients gives the effect of the algorithm relative to another procedure. The significance of this relative effect can be obtained using these vectors as basis for an F-test within the model.

### 2.10.3 Pairwise interactions

By the request of an anonymous reviewer we also investigate regressions with pairwise interactions. Note that the results from Table 2.5 are again to be taken with a grain of salt, as the models are significantly improved by adding further and higher interaction terms (see Appendix 2.10.1), which then also suggests the testing of hypotheses using the procedures from (names deleted to ensure anonymity).

The share of effect sizes smaller than 0.01 in absolute size for the pairwise interaction sales model is 65%, for the optimality model 71.67%. The share of p-values below 0.01 for the pairwise interaction sales model is 66.67% and for the optimality model 60%.

## 2.11 Appendix C: Supplementary material on generating functions

### 2.11.1 Albers' Original Table

The original table from Albers, reproduced as Table 2.6, has more columns and slightly different values. We start by explaining why we use fewer columns:

Albers compares different starting conditions with respect to the correlation of starting allocations with their optimal values. These starting conditions are constructed by changing an equal allocation, i.e. the budget is allocated evenly, each unit receives  $B/d = 8,000,000/8 = 1,000,000$  by roughly 5% towards a positive or negative correlation with the optimal allocation. However, this only makes sense in a situation without random disturbances. In our case, the magnitude of the error terms, even in the case of  $R^2 = 0.9$  (see Section 2.3.3), greatly exceeds a 5% boundary and hence varying starting conditions are not necessary. Therefore we set equal starting allocations with  $B/8$  for each unit.

Variable	Dependent Variable: Sales		Dependent Variable: Optimality	
	Coefficient	t-Value	Coefficient	t-Value
(Intercept)	0.778	139.155 ***	1.005	171.887 ***
Proc1	-0.001	-0.165	-0.003	-0.446
Proc2	0.017	2.781 **	-0.007	-1.016
Proc3	-0.006	-0.984	-0.010	-1.621
Form2	0.08	13.041 ***	-0.003	-0.451
Form3	0.072	11.721 ***	-0.046	-7.146***
Form4	-0.276	-44.992 ***	-0.113	-17.635***
Dist2	0.006	1.084	-0.004	-0.691
Dist3	0.012	2.141 *	-0.018	-3.04**
Elas1	0.043	8.904 ***	-0.047	-9.384***
Satu1	0.145	30.387 ***	0.012	2.484*
Budg8	0.043	9.049 ***	-0.026	-5.168***
Proc1:Form2	-0.007	-1.218	-0.007	-1.135
Proc2:Form2	0.047	8.068 ***	0.058	9.574***
Proc3:Form2	-0.001	-0.104	-0.005	-0.806
Proc1:Form3	0	0.071	-0	-0.082
Proc2:Form3	0.05	8.652 ***	0.058	9.528***
Proc3:Form3	0	-0.029	-0.002	-0.261
Proc1:Form4	-0.021	-3.634 ***	-0.028	-4.583***
Proc2:Form4	-0.076	-13.155 ***	-0.139	-23.085***
Proc3:Form4	-0.066	-11.412 ***	-0.081	-13.393***
Proc1:Dist2	0.002	0.327	0.002	0.476
Proc2:Dist2	-0.045	-9.045 ***	-0.049	-9.403***
Proc3:Dist2	0.003	0.515	0.006	1.182
Proc1:Dist3	0.001	0.154	-0.002	-0.332
Proc2:Dist3	-0.157	-31.487 ***	-0.187	-35.682***
Proc3:Dist3	0.008	1.596	0.014	2.717**
Proc1:Elas1	-0.014	-3.392 ***	-0.016	-3.666***
Proc2:Elas1	-0.044	-10.842 ***	-0.044	-10.197***
Proc3:Elas1	-0.005	-1.241	0.001	0.244
Proc1:Satu1	-0.013	-3.258 **	-0.014	-3.277**
Proc2:Satu1	-0.099	-24.192 ***	-0.1	-23.31***
Proc3:Satu1	-0.008	-1.976 *	-0.009	-2.088 *
Proc1:Budg8	0.004	0.873	0.006	1.296
Proc2:Budg8	0.036	8.869 ***	0.072	16.797***
Proc3:Budg8	0.029	7.138 ***	0.035	8.262***
Form2:Dist2	0.016	2.952 **	0.019	3.547***
Form3:Dist2	0.014	2.885 **	0.015	2.8**
Form4:Dist2	0.022	4.427 ***	-0.006	-1.084
Form2:Dist3	0.008	1.697	0.026	4.922***
Form3:Dist3	0.03	6.042 ***	0.041	7.834 ***
Form4:Dist3	0.05	9.954 ***	0	0.066
Form2:Elas1	-0.059	-14.342 ***	0.016	3.758***
Form3:Elas1	-0.011	-2.575 *	-0.024	-5.556 ***
Form4:Elas1	0.039	9.51 ***	0.037	8.579***
Form2:Satu1	-0.031	-7.576 ***	-0.017	-3.897***
Form3:Satu1	-0.097	-23.678 ***	0.031	7.166 ***
Form4:Satu1	0.009	2.283 *	-0.03	-7.051 ***
Form2:Budg8	-0.022	-5.282 ***	-0.001	-0.277
Form3:Budg8	0.064	15.614 ***	0.046	10.75***
Form4:Budg8	0.226	65.131 ***	0.155	36.357***
Dist2:Elas1	-0.007	-2.112 *	-0.011	-2.441*
Dist3:Elas1	-0.002	-0.444	-0.009	-0.792
Dist2:Satu1	-0.017	-4.685 ***	-0.003	-5.555***
Dist3:Satu1	-0.04	-11.201 ***	-0.021	-12.54***
Dist2:Budg8	-0.015	-4.113 ***	-0.046	-0.486
Dist3:Budg8	-0.029	-8.208 ***	0.002	1.216
Elas1:Satu1	0.021	7.205 ***	0.004	8.099 ***
Elas1:Budg8	-0.045	-15.479 ***	0	0.025
Satu1:Budg8	-0.020	-7.074 ***	-0.008	-2.683 **
$R^2 = 0.8157, R^2_{adj} = 0.8143$		$R^2 = 0.7245, R^2_{adj} = 0.7224$		
* p-value<0.05, ** p-value<0.01, *** p-value<0.001				

Table 2.5: Pairwise Interaction Effects Regression Models

No.	Products	Almost equal elasticities	Variated elasticities	Equal profit contribution at equally allocated budgets	Variated profit contribution at equally allocated budgets	Almost equal multipliers for saturation levels	Variated multipliers for saturation levels
1	A	0.26	0.11	2,500,000	1,000,000	6.1	3
2	B	0.27	0.12	2,500,000	1,000,000	6.2	10
3	C	0.28	0.13	2,500,000	4,000,000	6.3	3
4	D	0.29	0.14	2,500,000	4,000,000	6.4	10
5	E	0.31	0.47	2,500,000	1,000,000	6.6	3
6	F	0.32	0.48	2,500,000	1,000,000	6.7	10
7	G	0.33	0.49	2,500,000	4,000,000	6.8	3
8	H	0.34	0.50	2,500,000	4,000,000	6.9	10
	sum			20,000,000	20,000,000		

Table 2.6: Original Table from Albers

We shortly elaborate on the discrepancies in Albers (1997).

Firstly there is a problem concerning the degrees of freedom for determining the functions' parameters. e.g., for the modified exponential function which has two parameters,  $M_{exp}$  and  $h$ , Albers specifies three restrictions:

1. The elasticity at the optimal allocation must be the elasticity value from Table 2.6.
2. The contribution which is equivalent to sales in our model, as the contribution margins are never varied) at an input of  $x_i = 1,000,000$  must be the contribution value from Table 2.6.
3. The function's saturation must be 1,000,000 times the saturation multiplier from Table 2.6 .

This procedure leads to an overdetermined equation system. The ADBUDG functions have one additional parameter, but also one additional restriction ( $\phi_c < 1$  and  $\phi_S > 1$ ). That is why again overdetermined equation systems result.

Secondly, the third and seventh row from Table 2.6 suggest the existence of functions with a maximum of 3,000,000, but gives a contradictory value of 4,000,000 at an input of 1,000,000.

### 2.11.2 Visually testing S-shaped functions

We use visual control of the graphs of the S-shaped ADBUDG-functions to ensure that the function does not behave badly. The functions shown here are scaled such that they only take values between 0 and 1, where 1 is the actual saturation, and the exponent is greater than 1, i.e. there is an actual S-shape in the graph.

Functions with graphs such as the four in Figure 2.4 are rejected for the following reasons:

In the first image, the saturation is not reached within the domain of the function.

In the second image, the saturation is reached with an unreasonably small input.

In the third image, the lower bound 0 is not exceeded before a reasonable input.

In the fourth image, the non-constant part of the function is found on an unreasonably short interval of the input.

To make sure that our functions do not have such undesirable properties, we test them visually after generating parameter values.

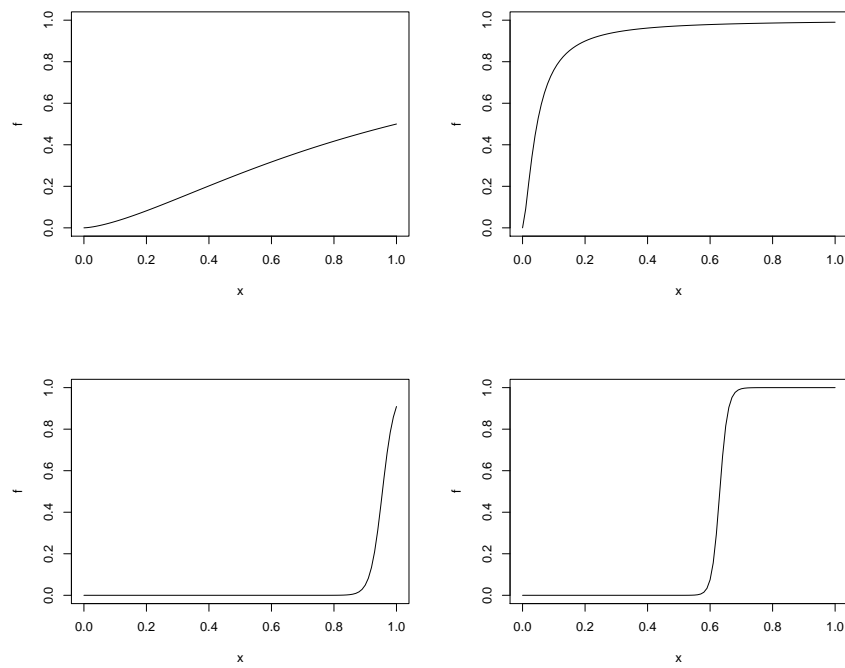


Figure 2.4: Undesirable function shapes

## Chapter 3

# Pricing Procedures

In this chapter, we compare pricing procedures in a monopolistic market over several situations. The article it was based on was accepted for review in the journal 'Business Research'. By the time of submitting this dissertation, it is still under review in that journal, but has not yet been published. The paper has two authors: Prof. Dr. Harald Hruschka and me. Changes compared to the original submission consist mainly of layout changes, and deletion of sections necessary only to the journals. The title of the paper as submitted to 'Business Research' was:

### **Monopolistic Pricing Rules Not Requiring Knowledge of the Price Response Function**

#### **ABSTRACT**

We conduct a simulation study to evaluate several heuristic rules which a decision maker might use to set the price of a product in a monopolistic market if neither the form nor the parameters of the underlying aggregate price response function are known. We consider several functional forms with different parameters and properties as well as error terms of different size. Rules are evaluated by the sum of forgone profits over planning horizons of different lengths. We group results based on different levels of elasticities, variable unit costs, and stochastic errors. Comparing the performance of the rules in each group enables a decision maker to choose the rule most appropriate for her/his situation. An extension of a rule proposed by Baumol and Quandt (1964) turns out best across all groups.

### 3.1 Introduction

We investigate several heuristic rules intended to determine price with respect to their capability to approximate optimal profit. Decision makers may resort to such rules if they know neither the form of the price response function nor its parameters. We evaluate rules by the sum of forgone profits over planning horizons of different lengths. Foregone profit in a period is defined as difference between the expected profit achieved by the rule and the expected optimal profit. Foregone profits are determined by a Monte Carlo study which generates unit sales (in the following briefly called sales) from static aggregate price response functions. Let us emphasize that the information about both form and parameters of the functions is only available for the Monte Carlo study, but remains hidden for the evaluated pricing rules.

A static aggregate price response function  $Q(p)$  links the price of a product to its sales. As prices of other products do not affect sales we have a monopolistic market. Such a price response function considers sales for an aggregate of costumers (e.g., all costumers residing in a certain region) and not sales for individual consumers. The function is static as it does not include dynamic effects (e.g., lagged prices or lagged sales) and its parameters are constant across time.

We assume throughout that the cost function is linear with known constant variable unit costs  $c$ . Therefore profit  $\Pi(p)$  as function of price  $p$  can be written as:

$$\Pi(p) = (p - c) Q(p) \quad (3.1)$$

We do not include fixed cost in expression (3.1) because it is irrelevant for the optimal solution. If the price response function and its parameter values are known, the profit-maximizing price  $p^*$  can be determined by solving the monopolistic pricing rule (also known as Amoroso-Robinson condition) which includes elasticities (Hruschka (1996); Hanssens et al (2001a); Hirschey et al (1993)):

$$p^* = \frac{\varepsilon}{\varepsilon + 1} c \quad (3.2)$$

For a differentiable response function price elasticities  $\varepsilon$  can be determined as point elasticities in the following way:

$$\varepsilon = \frac{\partial Q(p)}{\partial p} \frac{p}{Q(p)} \quad (3.3)$$

Except for the multiplicative response function (see 3.8 Appendix D) point elasticities vary with price.

Application of the monopolistic pricing rule (3.2) requires knowledge of the underlying price response function. This knowledge can be acquired by econometric methods, market experiments, customer surveys or expert interviews (Maurice et al (1992); Simon and Fassnacht (2016)). Econometric methods require at least a moderate number of sales and price observations. For pricing of new products this condition is never fulfilled. In addition observed prices should show sufficient variation. Market

experiments are an ideal alternative in terms of external validity, but decision makers usually avoid them because of very high costs. Customer surveys to obtain estimates of elasticities (e.g., by conjoint analysis) entail high costs. Expert interviews are less costly because of the low number of people involved. The external validity of both customer surveys and expert interviews may be low. In surveys, customers are exposed to an artificial situation, often with a focus on price. Therefore their responses may differ from those in the market place (East et al (2013)). It is well known that expert judgments as a rule perform worse than even very simple statistical models (Camerer and Johnson (1991)). In addition the assessment of price response of costumers by experts may be motivationally biased (e.g., sales managers tend to be overconfident and therefore underestimate the price sensitivity of customers (Markovitch et al (2015))).

As an alternative approach we investigate pricing rules which do not require that a decision maker knows the price response function and its parameters. These pricing rules are heuristic, which implies that they are not guaranteed to provide the optimal price. Nonetheless, profits generated by a rule should be as close as possible to profits which result at the unknown optimal price  $p^*$ .

Such pricing rules are especially appropriate if a decision maker faces a budget limitation and wants to set the price of a new product or the price of a product whose price variation in the past is low. Costs involved are much lower compared to experiments, customer surveys and expert interviews. Contrary to customer surveys and expert interviews these rules have high external validity as they process responses of customers in the market place.

We proceed as follows: In Section 3.2 we explain the details of the preparations of the simulation study, by presenting the decision problem and the overall gains function in Section 3.2.1 as well as the rules, response functions and error terms in Sections 3.2.2 to 3.2.10. In Section 3.3 we present the results of the simulation by performing comparison tests within certain groups outlined in Sections 3.3.1 and 3.3.2 and presented in Sections 3.3.3 to 3.3.5. We discuss the results and conclude with Section 3.4.

## 3.2 Preparing the Simulation

### 3.2.1 Decision Problem

Let  $Q: \mathbb{R}^+ \rightarrow \mathbb{R}^+$  be a static aggregate price response function, that maps the price in a period  $t$  to sales in the same period.  $Q$  itself is continuous, monotonically decreasing, and may have concave and convex areas, to allow for an S-shape. Furthermore,  $Q$  is not dependent on the period  $t$ .

In our simulation the price range is  $[1; 9]$ , the sales range  $[0; 1,000,000]$ . As we do not round off to integers, the entire simulation is scalable. In other words, these intervals can be transformed to different ranges and orders of magnitude yielding, as a rule, identical results.

In accordance with the marketing science literature we take into account that response functions are not deterministic (Hanssens et al (2001a); Leeflang et al (2015)). Therefore we add an error term to obtain a stochastic response function  $Q_s(p) = Q(p) + e$ . Errors are normally distributed  $e \sim \mathcal{N}(0, \sigma^2)$  for a given variance  $\sigma^2$  (whose construction will be explained in Section 3.2.9).

For a known price response function, it would be sufficient to consider expected sales profit because variable unit cost is constant (Jagpal (1999)). Expected profit can be written as

$$\Pi: \mathbb{R}^+ \rightarrow \mathbb{R}^+, p \mapsto (p - c) Q(p). \quad (3.4)$$

$\Pi(p)$  is continuous, and hence attains its maximum  $\Pi^*$  at a price  $p^*$  by Rudin (1976) (P.89/90 Theorem 4.16) since the interval  $[1; 9]$  is compact.

We define the *forgone profit*  $F(p)$  at a price  $p$  as difference between the expected profit  $\Pi(p)$  at a price  $p$  and the expected profit at the optimal price  $\Pi(p^*)$

$$F(p) := \Pi(p) - \Pi(p^*) \quad (3.5)$$

As  $\Pi(p^*)$  is the global maximum of  $\Pi$ ,  $F$  is non-positive, and  $F(p) = 0 \Leftrightarrow p = p^*$ . Therefore the global maximum of  $F$  is identical to the global maximum of  $\Pi$  and maximizing  $F$  is equivalent to maximizing  $\Pi$ . Furthermore, this maximum is unique for all price response functions which we consider in the simulation (see Section 3.2.4).

The decision problem which we investigate is different, because there is no information on the true price response function and its parameters. The decision maker only observes stochastic sales  $Q_s$  as responses to prices she or he has set in previous periods. Note that the optimal price, its corresponding profit and the expected profit for a price  $p$  are only determined during our Monte Carlo simulation, they are unknown to the decision maker and consequently not processed by any of the investigated rules.

Of course, the profit due to a price set by a rule should be as close as possible to the profit which results for the unknown optimal price  $p^*$ . Therefore we measure the performance of a pricing rule by the sum of forgone profits across a given planning



horizon of  $T$  periods

$$\sum_{t=1}^T F(p_t). \quad (3.6)$$

The higher this sum turns out for a pricing rule, the better we evaluate a rule.

With one exception, all the investigated rules require data of previous periods. Therefore stochastic sales and prices of the first two periods  $t = 1, 2$  are equal and deterministic for each rule. An equivalent perspective is that  $p_1, p_2, Q_s(p_1)$  and  $Q_s(p_2)$  are common knowledge. These values will be explained at the beginning of Section 3.3.

### 3.2.2 Pricing Rules

We compare eight pricing rules in total. Two of these rules are based on arc elasticities and the Amoroso-Robinson condition, two are based on approximating the price response function by a linear or multiplicative function and three are based on a publication of Thore (1964). For the purpose of the article, we assume that price response is elastic and deem elasticities outside the interval  $[-7; -1.25]$  unreasonable. Hence, if an estimated elasticity lies outside this interval, it is projected to the nearest boundary before determining the price  $p_t$  in period  $t$ .

The eight pricing rules are as follows:

1.) Rule based on arc elasticities (AE).

As the price response function is unknown, we estimate the arc elasticity  $\varepsilon_t$  from observed sales and prices of the previous two periods as:

$$\varepsilon_t = \frac{Q_s(p_{t-1}) - Q_s(p_{t-2})}{p_{t-1} - p_{t-2}} \frac{p_{t-2}}{Q_s(p_{t-2})} \quad (3.7)$$

(see Monroe (1999)) and insert it into the expression of the monopolistic pricing rule based on the Amoroso-Robinson condition (3.2) to determine the price in period  $t$ :

$$p_t^* = \frac{\varepsilon_t}{\varepsilon_t + 1} c \quad (3.8)$$

2.) Rule based on smoothed arc elasticities (SAE).

We apply first order exponential smoothing to reduce fluctuations of estimated elasticities. In period  $t = 3$  we define  $\tilde{\varepsilon}_3 := \varepsilon_3$  as in (3.7), and for  $t > 3$  we define

$$\tilde{\varepsilon}_t := \beta \tilde{\varepsilon}_{t-1} + (1 - \beta) \varepsilon_{t-1} \quad \text{with } \beta \in (0; 1) \quad (3.9)$$

and finally determine the price according to the Amoroso-Robinson condition (3.8) with  $\tilde{\varepsilon}_t$  replacing  $\varepsilon_t$ . A global optimization search provides  $\beta = 0.4$  as best value (see 3.5 Appendix A).

The next two rules approximate the unknown response function by either a multiplicative or a linear model using observed sales and prices up to period  $t$  for  $t = 3, 4, \dots, T$ . Note that usually coefficients of the approximating model change from period to period.

3.) Rule based on approximating by a multiplicative model (MM).

Log sales are approximated by estimating the double log response model  $\log(Q_s) = b_0 - b_1 \log(p)$  by OLS. Taking its antilog we obtain the multiplicative model  $Q_s = \exp(b_0)p^{-b_1}$ . If the multiplicative model is the true response function,  $-b_1$  equals the constant elasticity. As the response function is unknown, we consider  $-b_1$  to be an approximation to the unknown true elasticity. We define  $\varepsilon_t := -b_1$  and apply Equation (3.8) to compute  $p_t$ .

4.) Rule based on approximating by a linear model (LM).

We modify rule 3 presented in Baumol and Quandt (1964) which is based on estimating a linear function for sales and prices of only two periods. If the true price response function is stochastic, this approach easily runs into troubles giving a positive price coefficient. We approximate sales by a linear model  $Q_s = b_0 - b_1 p$  and compute

$$p_t = (b_0/b_1 + c)/2 \quad (3.10)$$

from Baumol and Quandt (1964). This expression is equivalent to the static monopolistic pricing rule if both price response and cost functions are linear and known.

The next three rules are based on a publication of Thore (1964) and look inter alia at observed profits  $\Pi_s(p)$  which are computed from observed sales by  $(p - c)Q_s(p)$ . Trying to find the profit maximizing solution they determine the price in period  $t$  as  $p_t = p_{t-1} + \Delta p_t$ . These rules differ from each other with respect to the computation of the price change  $\Delta p_t$ .

5.) Constant price changes (CC).

We extend Equation (4) in Thore (1964) which only outputs the sign of a price change by setting the absolute value of a price change to a constant value  $\alpha_1$ :

$$\Delta p_t = \alpha_1 \text{sign}((\Pi_s(p_{t-1}) - \Pi_s(p_{t-2}))(p_{t-1} - p_{t-2})). \quad (3.11)$$

The rule builds upon the idea that after a price increase which has led to higher (lower) profits the decision maker is inclined to increase (decrease) next period's price. According to the same reasoning, if a price decrease has led to lower (higher) profits, the decision maker increases (decreases) next period's price. This simple mechanism is reproduced by

$$\text{sign}((\Pi_s(p_{t-1}) - \Pi_s(p_{t-2}))(p_{t-1} - p_{t-2})). \quad (3.12)$$

If this expression equals one (minus one) price is increased (decreased) by the constant amount  $\alpha_1$ . A global optimization determines  $\alpha_1 = 0.3$  as best value (see 3.5 Appendix A).

6.) Price changes dependent on previous absolute price differences (DC).

This rule is similar to the previous rule, but allows price changes to vary. We use the modification of the rule given by Equation (5) in Thore (1964) defined as:

$$\Delta p_t = \alpha_2 \sqrt{|(p_{t-1} - p_{t-2})|} \text{sign}((\Pi_s(p_{t-1}) - \Pi_s(p_{t-2}))(p_{t-1} - p_{t-2})). \quad (3.13)$$

We decide to insert price differences under the square root instead of the profit differences of the original formulation to get similar orders of magnitude. An independent simulation confirms that our modification works better than the original rule from Thore (1964). Price changes are not constant because  $\alpha_2 \sqrt{|(p_{t-1} - p_{t-2})|}$  acts as stretch factor to the constant step size given by rule 5. This factor stabilizes the absolute amount of price changes, because higher (lower) previous price changes favour higher (low) current price changes in absolute terms.

The best value for  $\alpha_2$  is found to be 0.4 by the methods presented in 3.5 Appendix A.

7.) Slope dependent price changes (CS).

This rule is given by Equation (6) in Thore (1964):

$$\Delta p_t = \alpha_3 \frac{\Pi_s(p_{t-1}) - \Pi_s(p_{t-2})}{p_{t-1} - p_{t-2}}. \quad (3.14)$$

The rule is similar to gradient ascent of the profit function, but replaces the first derivative by the slope of the straight line connecting observed profits  $\Pi_s(p_{t-1})$  and  $\Pi_s(p_{t-2})$  (Chiang (1984)). For the price response functions investigated in the Monte Carlo Simulation gradient ascent with first derivative is known to find the the optimal solution. Alas, here the derivative cannot be computed because it requires knowledge of the price response function.

In our study slopes of observed profits take very high absolute values. We therefore decide to cushion the rule by the ratio of the interval centers  $p_{center}/Q_{center}$  ( $p_{center} = 5, Q_{center} = \text{Sat}/2$  with Sat as saturation level of the response function). Introducing this factor leaves us with

$$\Delta p_t = \alpha_3 (p_{t-1} - c) \frac{Q_s(p_{t-1}) - Q_s(p_{t-2})}{Q_{center}} \frac{p_{center}}{p_{t-1} - p_{t-2}}, \quad (3.15)$$

ensuring that  $\alpha_3$  has the same order of magnitude as  $\alpha_1$  and  $\alpha_2$ . The optimal value for  $\alpha_3$  (see 3.5 Appendix A) was found to be 0.1. Notice further that using this rule requires a vague knowledge of the function's saturation level.

8.) Random prices (RP).

Taken from Baumol and Quandt (1964), we sample the price from a uniform distribution whose lower bound equals unit cost  $c$  and whose upper bound equals 9.

Random prices serve as reference method in the evaluation, because the other pricing rules *should* perform better.

### 3.2.3 Parameter restrictions on Price Response Functions

We want to consider different functions with different properties. The properties of the five investigated functional forms, explained in Section 3.2.4, depend on their coefficients. We therefore start with a table of properties for the functions and choose coefficients such that the functions adhere to these properties.

We define a low, medium and high level for each of the three properties ‘Maximum’, ( $Q(1)$ , the highest possible sales value) ‘Minimum’ ( $Q(9)$ , the lowest possible sales value) and ‘Costs’. While the variable unit cost  $c$  itself does not influence the functional form, it is necessary for determining the optimal price  $p^*$  for profits and therefore the elasticities, which are defined as the point elasticities at the optimum, see Section 3.2.7. The values are taken from Table 3.1.

Table 3.1: Properties

Level	Maximum Sales	Minimum Sales	Variable Unit Costs
Variable name	$Max$	$Min$	$c$
low	500,000	100	2
medium	750,000	500	3
high	1,000,000	1,000	4

### 3.2.4 Functional Forms

We consider the following five widespread functional forms Hanssens et al (2001a):

- 1.) The linear function  $Q_{lin}(p) = a_0 - a_1 p$
- 2.) The multiplicative function  $Q_{mult}(p) = b_0 p^{-b_1}$
- 3.) The exponential function  $Q_{exp}(p) = e^{c_0 - c_1 p}$
- 4.) The semi-logarithmic function  $Q_{semlog}(p) = d_0 - d_1 \log(p)$
- 5.) The logistic function  $Q_{log}(p) = \frac{Q_{max}}{1 + e^{-(f_0 - f_1 p)}}$

with positive parameters  $a_0, a_1, b_0, b_1, c_0, c_1, d_0, d_1, f_0, f_1 > 0$  (see 3.6 Appendix B). Note that the logistic function has one additional degree of freedom with  $Q_{max}$  which is the asymptotic maximum of the function. For generating the coefficients, we always want two degrees of freedom and two restrictions, to be able to uniquely solve for the coefficients. Therefore, we add a third restriction for the logistic function, namely, that  $Q_{max}$  is 0.5% higher than the respective maximum sales value shown in Table 3.1 (i.e.,  $Q_{max} = Max \cdot 1.005$ ).

### 3.2.5 Setting Coefficients

As we want all of our functions to be comparable, we need them to have similar properties. We want the highest possible value to be  $Q(1) = \text{Max}$ , and the lowest possible value to be  $Q(9) = \text{Min}$ . With these two restrictions, we can uniquely solve for the coefficients of these functions (for a proof, see 3.6 Appendix B):

- 1.) for the linear function:

$$a_0 = \frac{9\text{Max}-\text{Min}}{8}, a_1 = \frac{\text{Max}-\text{Min}}{8}$$

- 2.) for the multiplicative function:

$$b_0 = \text{Max}, b_1 = \frac{\log(b_0/\text{Min})}{\log(9)}$$

- 3.) for the exponential function:

$$c_0 = \frac{9\log(\text{Max})-\log(\text{Min})}{8}, c_1 = \frac{\log(\text{Max})-\log(\text{Min})}{8}$$

- 4.) for the semi-logarithmic function:

$$d_0 = \text{Max}, d_1 = \frac{d_0-\text{Min}}{\log(9)}$$

- 5.) and for the logistic function:

$$\begin{aligned} Q_{\max} &= 1.005 \cdot \text{Max}, \\ f_0 &= \frac{9\log(\frac{\text{Max}}{Q_{\max}-\text{Max}})+\log(\frac{Q_{\max}-\text{Min}}{\text{Min}})}{8}, \\ f_1 &= \frac{\log(\frac{\text{Max}}{Q_{\max}-\text{Max}})+\log(\frac{Q_{\max}-\text{Min}}{\text{Min}})}{8}. \end{aligned}$$

### 3.2.6 Determining Optimal Prices

As explained in Section 3.2.1, we need to know the price maximizing the expected profit. So, for each function type, given its coefficients and variable unit cost  $c$  we need an exact formula for the optimal price  $p^*$ . For the semi-logarithmic and the logistic functions we need to solve equations of the shape  $y = x e^x$  for  $x$ , and do so by applying the Lambert- $W$ -function (see Corless et al (1996)), with  $x = W(y)$ . The derivation of the formulae, together with that of the well-definedness of the  $W$ -function can be found in 3.7 Appendix C, and they are as follows:

- 1.) for the linear function:

$$p^* = 1/2(\frac{a_0}{a_1} + c)$$

- 2.) for the multiplicative function:

$$p^* = \frac{-b_1 c}{-b_1 + 1}$$

this coincides with the Amoroso-Robinson condition (3.2) for a constant elasticity.

- 3.) For the exponential function:

$$p^* = \frac{1+c_1 c}{c_1}$$

- 4.) for the semi-logarithmic function:

$$p^* = \frac{d_1 c}{W(d_1 c e^{1-d_0})}$$

5.) and for the logistic function:

$$p^* = \frac{W(\exp(f_0 - 1 - cf_1)) + 1 + cf_1}{f_1}.$$

### 3.2.7 Evaluating Elasticities

To analyze the results of the Monte Carlo simulation we also need to know the point elasticities at the optimum price  $p^*$  (see Section 3.3.2) which for a differentiable price response function  $Q(p)$  are defined by expression (3.3). The point elasticities at  $p^*$  can be calculated as follows, with the derivation in 3.8 Appendix D:

1.) for the linear function:

$$\varepsilon = \frac{-a_1 p^*}{a_0 - a_1 p^*}$$

2.) for the multiplicative function:

$$\varepsilon = -b_1$$

3.) for the exponential function:

$$\varepsilon = -c_1 p^*$$

4.) for the semi-logarithmic function:

$$\varepsilon = \frac{-d_1}{d_0 - d_1 \log(p^*)}$$

5.) and for the logistic function:

$$\varepsilon = -f_1 p^* \left(1 - \frac{Q_{\log}(p^*)}{Q_{\max}}\right).$$

### 3.2.8 Restrictions between Rules and Functions

In order to ensure the reasonability of certain rules and functions, we need to pay attention to certain issues.

First of all, we do not allow for prices to be smaller than the unit cost  $c$  or larger than 9. Should that happen, the price is set to  $c$  or 9 respectively.

Furthermore, as we add a normally distributed error term to  $Q$  in order to obtain  $Q_s$ , it is possible to obtain non-positive sales. We prevent that, by setting an absolute minimum sales of 10 (an order of magnitude below the low minimum level), so the simulation uses the value  $\max(10, Q_s)$ .

This might lead to the fraction in the definition of rule 7.) ‘ThoreStep’ to be ill-defined, in which case  $\Delta p_t$  is randomly set to either  $+\alpha_3$  or  $-\alpha_3$ .

### 3.2.9 Error Terms

As mentioned before, we add a normally distributed error term  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$  to the deterministic part of the response function to obtain the stochastic response function  $Q_s$  as  $Q_s(p) := Q(p) + \varepsilon$ .

Again, we wish to be able to compare different functions while keeping the error level constant. We achieve that by means of the following procedure:

We choose a value for the coefficient of determination  $R^2$  from  $\{0.9, 0.7, 0.5\}$  corresponding to a low, medium and high error-term level, respectively. Then we perform an iterative search over values of  $\sigma^2$ . For each value of  $\sigma^2$  we input 2,000 prices uniformly distributed in the interval  $[1; 9]$  into the deterministic response function to obtain  $Q(p)$  to which we add errors from  $\mathcal{N}(0, \sigma^2)$  to obtain stochastic sales  $Q_s(p)$ . We then estimate the parameters of the response function by nonlinear least squares and compute its  $R^2$ -value. We stop the iterative search once the absolute difference to the desired  $R^2$ -value is less than 0.0001.

### 3.2.10 Iterations

We also wish to vary the number of iterations within the simulation. We chose a low, medium and high number of iterations of 10, 20 and 30 respectively. This *includes* the first two predetermined periods, to avoid a possible strictly linear relationship between the three levels.

## 3.3 Conducting the Simulation

Now that the preparations are complete, we can start conducting the simulation.

### 3.3.1 Experimental Designs

Each of the eight rules is used to perform an optimization for every possible constellation. The constellations result from combining 5 functional forms, 3 saturation levels, 3 minimum levels, 3 variable unit cost levels, 3 error term levels, and 3 iteration levels. We therefore look at a total of  $5 \times 3^5 = 1,215$  constellations.

As most of the rules depend on previous values, we start with the following values for the first two periods: The price in period 1 is  $1.5 \cdot c$ , i.e. 50% above the cost, and the price in period 2 is  $\frac{e \cdot c}{1+e}$  where  $c$  is the cost and  $e$  is a meta-elasticity of -2.5. In other words, the price in period 2 is the Amoroso-Robinson condition for an elasticity of -2.5.

### 3.3.2 Performing Tests

The aim of this article is to give advice dependent on the situation the marketing manager faces. We therefore split our evaluation based on values of the following three descriptors of the decision situation:

- 1.) price elasticity,
- 2.) variable unit cost, and
- 3.) size of the error terms.

When preparing the simulation, we use variable costs and error terms as factors in the experimental design, i.e., we are able to choose them before generating the functions. On the other hand, point elasticities at the optimal price are *implied* by these functions. Therefore the cost and error term factors are balanced in the experimental design,

whereas elasticities are not. As elasticities are on a metric scale we group them into the intervals  $[-6; -4]$ ,  $(-4; -3]$ ,  $(-3; -2]$ ,  $(-2; -1]$  leading to relative group sizes of 5%, 16%, 47% and 31%, respectively. We note that this distribution of elasticities is very similar to the one documented in a relevant meta-analysis (Bijmolt et al (2005)) with relative group sizes of 16%, 19%, 32% and 31%. A Bhattacharyya coefficient of 0.96 provides evidence to this high similarity (Aherne et al (1997)).

### 3.3.3 Recommended Rules

The comparisons within a certain group are conducted as follows:

We take the subset of data that belongs to the group we are considering (e.g., all functions with a point elasticity within  $(-3, -2]$  at the optimal price). We then award points to the rules in the following manner: For each pair of rules  $(A, B)$ , we perform a Tukey-test within the ANOVA of that subset. If the p-value of the Tukey-Test is higher than 0.05, we ignore it. If it is lower, we look at the coefficient of the Tukey-Test (i.e., the difference of effects  $A - B$ ). If it is positive, then the average foregone profits (a negative number) of  $B$  is lower (i.e., has a higher absolute value) than those of  $A$ . This means that  $A$  yielded profits closer to the optimum, and  $A$  is awarded a point. Otherwise it is negative, and  $B$  is awarded a point.

After analyzing the scores within each group, we add them up to get a total score for the entire grouping. *Note that these total scores differ dependent on the grouping, since the points are only awarded for **significant** differences in effects.*

Table 3.2: Recommended Rules

Elasticity /Cost/ $R^2$	RP	LM	MM	AE	SAE	CC	CS	DC
$\varepsilon \in (-2; -1]$	3	5	0	0	2	5	3	5
$\varepsilon \in (-3; -2]$	3	5	1	0	1	6	3	6
$\varepsilon \in (-4; -3]$	0	1	1	0	1	1	1	3
$\varepsilon \in [-6; -4]$	0	0	2	0	0	0	0	0
Total	6	11	4	0	4	12	7	14
$c = 2$	3	5	0	0	2	5	3	5
$c = 3$	3	5	1	0	1	5	3	6
$c = 4$	3	5	1	0	1	5	1	5
Total	9	15	2	0	4	15	7	16
$R^2 = 0.9$	3	5	0	0	1	5	3	5
$R^2 = 0.7$	3	5	0	0	1	5	3	5
$R^2 = 0.5$	3	4	0	0	2	5	2	5
Total	9	14	0	0	4	15	8	15

Overall, two main conclusions can be drawn from the top third of Table 3.2, which groups by point elasticities: First and foremost, the best rules overall are ‘LM’, ‘CC’ and ‘DC’, as they are superior to most other rules in most circumstances. However when the elasticity lies in the interval of  $[-6; -4]$ , all rules are more or less equivalent. Only ‘MM’ is significantly superior to other rules. In fact, when looking at the actual



Tukey-Tests, it is superior precisely to ‘CC’ and ‘DC’ (see also the top third of Table 3.3).

We therefore conclude: In the area of moderate elasticities (i.e.,  $(-4; -1]$ ) ‘LM’ as well as ‘CC’ and ‘DC’ are advisable. Remember however, that the latter two require determining a step factor. Furthermore, in an area of extreme elasticities (i.e.,  $[-6; -4]$ ) ‘MM’ is to be preferred.

The middle section of Table 3.2, which groups by costs, looks quite similar to the top third. First of all, the advisable rules are once again ‘LM’, ‘CC’ and ‘DC’. However this time, this comes without the caveat of extreme situations, i.e., none of the groups behaves as an outlier.

Therefore, we similarly recommend using those three rules and remark the necessity of a step constant for ‘CC’ and ‘DC’.

The conclusions to be drawn from the bottom third of Table 3.2 are equivalent to those from the middle.

### 3.3.4 Unrecommended Rules

By means of essentially the same method, we can advise *against* some of the rules. Instead of giving a score of 1 for every time a rule is superior to another, we give a score of -1 for every time a rule is *inferior* to another one at a p-value of the Tukey-Test of at most 0.05.

Elasticity /Cost/ $R^2$	RP	LM	MM	AE	SAE	CC	CS	DC
$\varepsilon \in (-2; -1]$	-3	0	-6	-6	-5	0	-3	0
$\varepsilon \in (-3; -2]$	-3	-2	-5	-7	-5	0	-3	0
$\varepsilon \in (-4; -3]$	-1	0	0	-6	0	0	-1	0
$\varepsilon \in [-6; -4]$	0	0	0	0	0	-1	0	-1
Total	-7	-2	-11	-19	-10	-1	-7	-1
$c = 2$	-3	0	-6	-6	-5	0	-3	0
$c = 3$	-3	-1	-5	-7	-5	0	-3	0
$c = 4$	-3	0	-4	-7	-4	0	-3	0
Total	-9	-1	-15	-20	-14	0	-9	0
$R^2 = 0.9$	-3	0	-5	-6	-5	0	-3	0
$R^2 = 0.7$	-3	0	-5	-6	-5	0	-3	0
$R^2 = 0.5$	-2	0	-6	-6	-4	0	-3	0
Total	-8	0	-16	-18	-14	0	-9	0

As suspected from the previous sections, the top third of Table 3.3 confirms that the rules based on Amoroso-Robinson (‘AE’ and ‘SAE’) are not advisable. For the interval  $(-3; -1]$  we see that additionally ‘MM’ is very inadvisable, in stark contrast

to  $[-6; -4]$ , where it is the only recommendable rule. It comes to no surprise that randomly selecting a price, i.e., rule 'RP' is hardly ever a good idea.

The middle section of Table 3.3 further confirms the results obtained so far. The Amoroso-Robinson-based rules, as well as 'RP' and 'CS' are inadvisable. So is 'MM', especially when the cost level is comparatively low.

Finally, the bottom third of Table 3.3 rounds out what we have seen before: Neither the Amoroso-Robinson rules, nor 'RP', 'MM' or 'CS' are generally advisable.

### 3.3.5 Robustness

We consider results of the the Tukey-Test as significant only if the resulting  $p$ -value was below a threshold of  $\alpha = 5\%$ . While this may be a common level for significance, changing this value might change the results. We therefore performed the same analyzes with alternative  $\alpha$ -values of 10%, 1% and 0.5% and the results were almost unchanged. This is due to the fact that the vast majority of significant comparisons were so at a level far below 0.1%. We therefore conclude that our method is robust with respect to the threshold  $\alpha$ .

## 3.4 Conclusion

We conduct a simulation study to evaluate rules of thumb for setting the price in a monopolistic market for one product with respect to their capability to approximate optimal profit. Decision makers may resort to such rules if they know neither the form of the price response function nor its parameters.

Summarizing what we have seen in the Tables 3.2 and 3.3, we can as a rule recommend the rules 'LM', 'CC' and 'DC', however 'CC' and 'DC' require setting an additional constant. In the special case of extreme elasticities, only 'MM' is advised. However, usually 'MM' is inferior to other rules, and is only advised in this case and possibly in the situations with high cost levels. It does not supersede 'LM'. In general 'LM' is the most advisable rule overall. Table 3.3 further yields that the rules 'AE', 'SAE', 'CS' and 'RP' are inadvisable, as well as 'MM', except for very few situations.

We try to encompass as many different situations as possible to ensure that results are applicable by marketing managers in a flexible way. Results are not only surprisingly constant, they also yield a clear favorite, when optimizing and statistically analyzing forgone profits. This favorite is a modification of a rule introduced in Baumol and Quandt (1964). This rule is based on a mathematical derivation of the optimal price given a linear price response function and a linear cost function. The rule also uses linear regression, which is able to handle different additive error terms. These properties cause the rule to be the best performing overall in Section 3.3.3, and rarely being unrecommended in Section 3.3.4.

### 3.5 Appendix A: Optimization of Parameters in Rules of Thumb

To find an optimal value for a parameter of a certain rule, we proceed as follows:

For a parameter  $\mu \in \mathbb{R}$  and a rule  $R = R(\mu)$  which depends on the parameter  $\mu$ , let  $p_t(R(\mu))$  be the price determined by rule  $R$  given parameter  $\mu$ . For  $t = 1, 2$  this is the common-knowledge price from the first two periods, and for  $t > 2$  it is whatever rule  $R$  calculates, based on previous periods' data. We have four rules that depend on such a parameter:

- 1.) Rule 2 (R=SAE) needs a smoothing constant  $\mu = \beta$ ,
- 2.) Rule 5 (R=CC) needs a constant step size  $\mu = \alpha_1$ ,
- 3.) Rule 6 (R=DC) needs a step size factor  $\mu = \alpha_2$ , and
- 4.) Rule 7 (R=CS) likewise needs a step size factor  $\mu = \alpha_3$ .

We then determine a reasonable range for  $\mu$  and then discretize this range to a number of fixed values  $\mu_1, \dots, \mu_9$ .

For R=SAE, the range for  $\mu = \beta$  is already  $(0; 1)$ , and we choose the nine discrete values  $\mu_1 = 0.1, \mu_2 = 0.2, \dots, \mu_9 = 0.9$  with a distance of 0.1 between consecutive values.

For R=CC and R=DC respectively, the most useful range for  $\mu = \alpha_1$  and  $\mu = \alpha_2$  is  $(0; 1)$  as well, so we also choose the same discrete values.

Based on further analysis, for R=CS any value larger than 0.5 was useless, so from the range  $(0; 0.5)$  we choose the values  $\mu_1 = 0.05, \mu_2 = 0.1, \dots, \mu_8 = 0.4, \mu_9 = 0.45$  with a distance of 0.05 between consecutive values.

Given a rule  $R$ , for each possible index  $i = 1, \dots, 9$  we perform the price finding optimization from the simulation study for the parameter  $\mu_i$  by setting the price to be  $p_t(R(\mu_i))$  for period  $t$  and obtain an average performance value of (forgone) profits and proximity to the optimal price. Based on the rankings of those two performances compared among all values  $\mu_1, \dots, \mu_9$ , we chose the parameter  $\mu_i$  with the overall best ranking.

We obtained the values of  $\beta = 0.4$  for R=SAE,  $\alpha_1 = 0.3, \alpha_2 = 0.4, \alpha_3 = 0.1$ , for R=CC, DC and CS respectively.

For R=SAE, we repeated the process with a finer step-size of 0.05, which led to the same result

### 3.6 Appendix B: Derivations of Coefficients

In this section, we derive the coefficients of the five functions, given their properties. As mentioned earlier, we want all deterministic functions to achieve their predetermined maximum value  $Sat$  at a price of  $p = 1$ , and the predetermined minimal value  $Min$  at  $p = 9$ . For purpose of readability, we call the maximum  $Sat$  instead of  $Max$  throughout this Appendix. Remember that our functions have two parameters each, leading to two degrees of freedom, with one exception: The logistic function has a third parameter  $Q_{max}$ , which is the asymptotic maximum, or supremum. So we set  $Q_{max} := 1.005 \cdot Sat$ , so the supremum is 0.5% above the value at  $p = 1$ . All we need to do is start with the equations  $Q(1) = Sat, Q(9) = Min$  and possibly  $Q_{max} := 1.005 \cdot Sat$  and solve for the parameters of the function. We furthermore show that all coefficients are indeed positive. At some point we use the fact, that the maximal attained value of  $Min$  is 1,000, and the minimal attained value of  $Sat$  is 500,000.

- 1.) The linear function  $Q_{lin}(p) = a_0 - a_1 \cdot p$ .

We start from

$$Q(1) = a_0 - a_1 = Sat \quad (3.6.1)$$

$$Q(9) = a_0 - 9a_1 = Min. \quad (3.6.2)$$

This is just an ordinary system of 2 linear equations, so define

$$\mathbf{M} := \begin{pmatrix} 1 & -1 \\ 1 & -9 \end{pmatrix}, \mathbf{a} := \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}, \mathbf{V} := \begin{pmatrix} Sat \\ Min \end{pmatrix}, \quad (3.6.3)$$

and all that's left is  $\mathbf{Ma} = \mathbf{V}$ , with  $\mathbf{M}^{-1} := \begin{pmatrix} 9/8 & -1/8 \\ 1/8 & -1/8 \end{pmatrix}$  and hence  $\mathbf{a} = \mathbf{M}^{-1}\mathbf{V}$ , in particular

$$a_0 = \frac{9Sat - Min}{8}, \quad (3.6.4)$$

$$a_1 = \frac{Sat - Min}{8}, \quad (3.6.5)$$

both of which are positive, as  $Sat > Min$ .

- 2.) The multiplicative function  $Q_{mult}(p) = b_0 \cdot p^{-b_1}$ .

The step  $Q(1) = b_0 = Sat > 0$  is fairly straightforward, and then

$$Q(9) = Sat \cdot 9^{-b_1} = Min \quad (3.6.6)$$

$$\Rightarrow -b_1 \log(9) = \log(Min/Sat) \quad (3.6.7)$$

$$\Rightarrow b_1 = \frac{\log(Sat/Min)}{\log(9)}. \quad (3.6.8)$$

Again,  $b_1 > 0$  as  $Sat > Min$ .

3.) The exponential function  $Q_{exp}(p) = e^{c_0 - c_1 \cdot p}$ .

Taking the logarithm on both sides of the equations  $Q(1) = Sat$  and  $Q(9) = Min$  again leads to a system of linear equations with the same matrix  $\mathbf{M}$  as in (3.6.3). We can therefore proceed as in the linear case to obtain:

$$c_0 = \frac{9 \log(Sat) - \log(Min)}{8} \quad (3.6.9)$$

$$c_1 = \frac{\log(Sat) - \log(Min)}{8} = \frac{\log(Sat/Min)}{8}. \quad (3.6.10)$$

Again, since  $Sat > Min$ , both  $c_0$  and  $c_1$  are positive.

4.) The semi-logarithmic function  $Q_{semilog}(p) = d_0 - d_1 \cdot \log(p)$ .

Again,  $Q(1) = d_0 = Sat > 0$  is straightforward, followed by (3.6.11)

$$Q(9) = Sat - d_1 \log(9) = Min \quad (3.6.12)$$

$$\Rightarrow d_1 \log(9) = Sat - Min \quad (3.6.13)$$

$$\Rightarrow d_1 = \frac{Sat - Min}{\log(9)} > 0. \quad (3.6.14)$$

5.) The logistic function  $Q_{log}(p) = \frac{Q_{max}}{1 + e^{-(f_0 - f_1 \cdot p)}}$ .

The equations

$$Q(1) = Sat = \frac{Q_{max}}{1 + \exp(f_1 - f_0)}, \text{ and} \quad (3.6.15)$$

$$Q(9) = Min = \frac{Q_{max}}{1 + \exp(9f_1 - f_0)} \quad (3.6.16)$$

can be transformed into

$$f_1 - f_0 = \log\left(\frac{Q_{max} - Sat}{Sat}\right) \quad (3.6.17)$$

$$9f_1 - f_0 = \log\left(\frac{Q_{max} - Min}{Min}\right). \quad (3.6.18)$$

We rename the right hand side of (3.6.17) and (3.6.18) as  $\hat{S} := \log\left(\frac{Q_{max} - Sat}{Sat}\right)$  and  $\hat{M} := \log\left(\frac{Q_{max} - Min}{Min}\right)$ . and again summarize the equations as

$$-\mathbf{M} \begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = \begin{pmatrix} \hat{S} \\ \hat{M} \end{pmatrix} \quad (3.6.19)$$

and solve my multiplying with  $-\mathbf{M}^{-1}$  from the left to obtain

$$f_0 = \frac{-9\hat{S} + \hat{M}}{8} \quad (3.6.20)$$

$$f_1 = \frac{-\hat{S} + \hat{M}}{8}. \quad (3.6.21)$$

It remains to show that both coefficients are positive. Note that

$$\hat{M} = \log\left(\frac{Q_{\max} - \text{Min}}{\text{Min}}\right) = \log(Q_{\max} - \text{Min}) - \log(\text{Min}) \quad (3.6.22)$$

and since

$$\text{Min} \leq 1000, Q_{\max} > S \geq 500.000 \quad (3.6.23)$$

we have

$$Q_{\max} - \text{Min} > \text{Sat} - \text{Min} \geq 500.000 - 1000 = 499.000 \quad (3.6.24)$$

and hence

$$\hat{M} = \log(Q_{\max} - \text{Min}) - \log(\text{Min}) > \log(499.000) - \log(1000) > 6. \quad (3.6.25)$$

Furthermore

$$\hat{S} = \log\left(\frac{Q_{\max} - \text{Sat}}{\text{Sat}}\right) = \log\left(\frac{1.005\text{Sat} - \text{Sat}}{\text{Sat}}\right) = \log(0.005) < -5, \quad (3.6.26)$$

and hence

$$f_0 = \frac{-9\hat{S} + \hat{M}}{8} \geq 51/8 > 0 \quad (3.6.27)$$

$$f_1 = \frac{-\hat{S} + \hat{M}}{8} \geq 11/8 > 0. \quad (3.6.28)$$

### 3.7 Appendix C: Derivations of Prices

In this section we determine the optimal price within the price range, and show that it is indeed optimal.

We define the profit-function as in (3.1) as  $\Pi(p) = (p - c) \cdot Q(p)$  and solve  $0 = \frac{\partial \Pi}{\partial p} = Q(p) + (p - c) \frac{\partial Q}{\partial p}$  for  $p$ .

This always leads to a *unique* solution  $p^*$  in the case of our five functions. We show that  $\frac{\partial^2 \Pi}{\partial p^2} = 2 \cdot \frac{\partial Q}{\partial p} + (p - c) \frac{\partial^2 Q}{\partial p^2}$  is negative, as either

- 1.) most sales functions  $Q$  are decreasing with negative curvature, i.e., the first two derivatives are non-positive, or
- 2.) we show directly that  $\frac{\partial^2 \Pi}{\partial p^2} < 0$  everywhere, or
- 3.) we show that  $\frac{\partial^2 \Pi}{\partial p^2} < 0$  at  $p^*$ .

By Lemma 3.7.1 below, we conclude that even in the case of 3.), the optimum is a global maximum.

**Lemma 3.7.1.** *In case 3.), the local optimum must be a global one as well.*

*Proof.* We assume that the opposite holds and derive a contradiction. Assume, there is a point  $p'$  with  $\Pi(p') > \Pi(p^*)$  (which is necessarily a boundary point, but that is irrelevant).

Without loss of generality, let  $p' < p^*$ .

As  $p^*$  is a local maximum,  $\exists \varepsilon > 0, p_0 < p^*$  such that  $p^* - p_0 < \varepsilon, \Pi(p^*) - \Pi(p_0) > 0$  and  $p_0 > p'$ , in other words, there is a point  $p_0$  between  $p'$  and  $p^*$  with a smaller profit value than  $p^*$ .

By the intermediate value theorem (Theorem 4.23 in Rudin (1976)), we know there is another value  $p_1$  between  $p'$  and  $p_0$  such that  $\Pi(p_1) = \Pi(p^*)$ .

By the mean value theorem (Theorem 5.10 in Rudin (1976), applied with  $f(a) = f(b)$ , also known as Rolle's Theorem), there must be yet another value  $p_2$  between  $p_1$  and  $p^*$  such that  $\frac{\partial \Pi}{\partial p}(p_2) = 0$

That is a contradiction to  $p^*$  being the unique solution to  $\frac{\partial \Pi}{\partial p}(p) = 0$

□

This lemma makes use of two facts about our function:

Firstly, all our functions are twice differentiable on their range. This is well known, as they are compositions of smooth functions with no singularities near their range.

Secondly, the uniqueness of  $p^*$ . We will indeed see below, that for each function considered,  $\frac{\partial \Pi}{\partial p}(p) = 0$  has exactly one solution.

Furthermore, we make use of the Lambert W function, see Corless et al (1996) for a discussion.

1.) The linear function  $Q_{lin}(p) = a_0 - a_1 \cdot p$ .

$Q_{lin}$  has derivatives  $\frac{\partial Q}{\partial p} = -a_1 < 0$  and  $\frac{\partial^2 Q}{\partial p^2} = 0$ . Hence

$$0 = Q(p) + (p - c) \frac{\partial Q}{\partial p} \quad (3.7.1)$$

$$= a_0 - a_1 \cdot p + (p - c) \cdot (-a_1) \quad (3.7.2)$$

$$= -2 \cdot a_1 p + a_1 c + a_0 \quad (3.7.3)$$

$$\Rightarrow p^* = \frac{a_1 \cdot c + a_0}{2a_1} \quad (3.7.4)$$

2.) The multiplicative function  $Q_{mult}(p) = b_0 \cdot p^{-b_1}$ .

$Q_{mult}$  has first and second derivatives  $\frac{\partial Q}{\partial p} = -b_1 b_0 p^{-b_1-1} < 0$  and  $\frac{\partial^2 Q}{\partial p^2} = (-b_1 - 1)(-b_1) b_0 p^{-b_1-2} > 0$ .

Nevertheless,

$$\frac{\partial^2 \Pi}{\partial p^2} = 2 \cdot b_1 b_0 p^{b_1-1} + (p - c) \cdot (b_1 - 1) b_1 b_0 p^{b_1-2} \quad (3.7.5)$$

$$= b_0 b_1 (2p^{b_1-1} + (b_1 - 1)p^{b_1-1} - c(b_1 - 1)p^{b_1-2}) \quad (3.7.6)$$

$$= b_0 b_1 p^{b_1-1} (2 + (b_1 - 1)(1 - c/p)) < 0 \quad (3.7.7)$$

$$\Leftrightarrow (2 + (b_1 - 1)(1 - c/p)) > 0 \quad (3.7.8)$$

$$\begin{aligned} p^* = c \frac{-b_1}{1-b_1} \\ \Leftrightarrow 2 + (b_1 - 1) \left(1 - \frac{b_1 + 1}{b_1}\right) > 0 \end{aligned} \quad (3.7.9)$$

$$\Leftrightarrow 2 + (b_1 - 1) \left(\frac{b_1 - b_1 - 1}{b_1}\right) = 2 + \frac{1 - b_1}{b_1} > 0 \quad (3.7.10)$$

$$\Leftrightarrow 2b_1 - b_1 + 1 < 0 \quad (3.7.11)$$

$$\Leftrightarrow b_1 < -1. \quad (3.7.12)$$

As we have seen in the section about the distribution of the elasticities,  $b_1 < -1$  is always satisfied. Remains to show that indeed  $p^* = c \frac{b_1}{1+b_1}$ :

$$0 = Q(p) + (p - c) \frac{\partial Q}{\partial p} = b_0 \cdot p^{b_1} + (p - c) b_1 b_0 p^{b_1-1} \quad (3.7.13)$$

$$\Rightarrow p + b_1 p - b_1 c = 0 \quad (3.7.14)$$

$$\Rightarrow p(1 + b_1) = cb_1 \quad (3.7.15)$$

$$\Rightarrow p = c \frac{b_1}{1 + b_1} \quad (3.7.16)$$

3.) The exponential function  $Q_{exp}(p) = e^{c_0 - c_1 \cdot p}$ .



$Q_{exp}$  has derivatives  $\frac{\partial Q}{\partial p} = Q(p) \cdot (-c_1) < 0$  and  $\frac{\partial^2 Q}{\partial p^2} = Q(p) \cdot c_1^2 > 0$ . Nevertheless,

$$\frac{\partial^2 \Pi}{\partial p^2} = 2 \cdot Q(p) \cdot (-c_1) + (p - c)Q(p) \cdot c_1^2 < 0 \quad (3.7.17)$$

$$\Leftrightarrow -2 \cdot c_1 + (p - c)c_1^2 < 0 \quad (3.7.18)$$

$$p^* = \frac{1+cc_1}{c_1} \Leftrightarrow -2 \cdot c_1 + c_1(1 + cc_1) - c \cdot c_1^2 < 0 \quad (3.7.19)$$

$$\Leftrightarrow -2 \cdot c_1 + c_1 + cc_1^2 - cc_1^2 < 0 \quad (3.7.20)$$

$$\Leftrightarrow c_1 > 0 \quad (3.7.21)$$

Remains to show that indeed  $p^* = \frac{1+cc_1}{c_1}$ :

$$0 = Q(p) + (p - c) \frac{\partial Q}{\partial p} = Q(p) + (p - c) \cdot (-c_1) \cdot Q(p) \quad (3.7.22)$$

$$\Leftrightarrow 1 - (p - c)c_1 = 0 \quad (3.7.23)$$

$$\Leftrightarrow p - c = 1/(c_1) \quad (3.7.24)$$

$$\Leftrightarrow p = \frac{1 + cc_1}{c_1} \quad (3.7.25)$$

4.) The semi-logarithmic function  $Q_{semlog}(p) = d_0 - d_1 \cdot \log(p)$ .

$Q_{semlog}$  has derivatives  $\frac{\partial Q}{\partial p} = -d_1/p < 0$  and  $\frac{\partial^2 Q}{\partial p^2} = d_1/p^2 > 0$ , and yet

$$\frac{\partial^2 \Pi}{\partial p^2} = 2 \cdot -d_1/p + (p - c) \cdot d_1/p^2 < 0 \quad (3.7.26)$$

$$\Leftrightarrow -2d_1 + d_1 - c \cdot d_1/p = -d_1(1 + c/p) < 0, \quad (3.7.27)$$

and we wish to find the optimal price. We begin by solving for the logarithm

$$0 = \frac{\partial \Pi}{\partial p} = Q(p) + (p - c) \frac{\partial Q}{\partial p} = d_0 - d_1 \cdot \log(p) + (p - c) \cdot (-d_1/p) \quad (3.7.28)$$

$$\Rightarrow d_1 \cdot \log(p) = d_0 + (p - c)(-d_1/p) = d_0 - d_1 + cd_1/p \quad (3.7.29)$$

$$\Rightarrow \log(p) = d_0/d_1 - 1 + c/p, \quad (3.7.30)$$

Then, exponentiation, and molding it into a function of the shape  $ye^y$  to apply the Lambert-W-Function:

$$\Rightarrow p = \exp(d_0/d_1 - 1) \exp(c/p) \quad (3.7.31)$$

$$\Rightarrow c \exp(1 - d_0/d_1) = c/p \exp(c/p) \quad (3.7.32)$$

$$\Rightarrow W(c \exp(1 - d_0/d_1)) = W(c/p \exp(c/p)) = c/p \quad (3.7.33)$$

Finally, we can solve for  $p$ :

$$p = c/W(c \exp(1 - d_0/d_1)). \quad (3.7.34)$$

$W(y)$  is always a real, positive number, as long as  $y > 0$ . So for well-definedness, we check that  $c \exp(1 - d_0/d_1) > 0$  which is trivial.

5.) The logistic function  $Q_{log}(p) = \frac{Q_{max}}{1 + e^{-(f_0 - f_1 \cdot p)}}$ .

We define  $D(p) := 1 + e^{-(f_0 - f_1 \cdot p)}$  as the denominator and  $E(p) := D(p) - 1 = e^{-(f_0 - f_1 \cdot p)}$  as the exponential part of the function, to make calculations shorter and easier. In particular, we have  $\frac{\partial D(p)}{\partial p} = \frac{\partial E(p)}{\partial p} = E(p) \cdot f_1$ .

$Q_{log}$  has derivatives

$$\frac{\partial Q}{\partial p} = \frac{-Q_{max}}{D(p)^2} \exp(-(f_0 - f_1 \cdot p)) f_1 < 0 \text{ and} \quad (3.7.35)$$

$$\frac{\partial^2 Q}{\partial p^2} = \frac{2Q_{max}}{D(p)^3} [E(p) f_1]^2 + \frac{-Q_{max}}{D(p)^2} E(p) f_1^2 \quad (3.7.36)$$

$$= \frac{Q_{max}}{D(p)^2} E(p) f_1^2 (2E(p)/D(p) - 1) \quad (3.7.37)$$

For the sign of the second derivative we use the fact that at the optimum  $p^*$  we have  $D(p) = (p - c)f_1 E(p)$ , which is shown further down in (3.7.49). The second derivative is

$$\frac{\partial^2 \Pi}{\partial p^2} = 2 \frac{\partial Q}{\partial p} + (p - c) \frac{\partial^2 Q}{\partial p^2} \quad (3.7.38)$$

$$= \frac{-2Q_{max}E(p)f_1}{D(p)^2} E(p)f_1 + (p - c) \frac{Q_{max}}{D(p)^2} E(p)f_1^2 \left( \frac{2E(p)}{D(p)} - 1 \right) \quad (3.7.39)$$

$$= \frac{Q_{max}E(p)f_1}{D(p)^2} (-2 + (p - c)f_1 \left( \frac{2E(p)}{D(p)} - 1 \right)) \quad (3.7.40)$$

$$= \frac{Q_{max}E(p)f_1}{D(p)^3} (-2D(p) + (p - c)f_1(2E(p) - D(p))) \quad (3.7.41)$$

$$= \frac{Q_{max}E(p)f_1}{D(p)^3} (-2D(p) + (p - c)f_1(E(p) - 1)) \quad (3.7.42)$$

$$\stackrel{(3.7.49)}{=} \frac{Q_{max}E(p)f_1}{D(p)^3} (-2D(p) + D - (p - c)f_1) \quad (3.7.43)$$

$$= \frac{Q_{max}E(p)f_1}{D(p)^3} (-D(p) - (p - c)f_1), \quad (3.7.44)$$

which is indeed negative. So we need to show that  $p^*$  indeed has the claimed shape, as well as (3.7.49).

Start from

$$0 = \frac{\partial \Pi}{\partial p} = Q(p) + (p - c) \frac{\partial Q}{\partial p} \quad (3.7.45)$$

$$= Q(p) + (p - c) \frac{-Q_{max}}{D(p)^2} E(p) \cdot f_1 \quad (3.7.46)$$

$$= Q(p) + Q(p)(p - c) \frac{-1}{D(p)} E(p) \cdot f_1 \quad (3.7.47)$$

$$= Q(p)(1 + (p - c) \frac{-f_1}{D(p)} E(p)) \quad (3.7.48)$$

which is equivalent to the desired

$$D(p) = (p - c)f_1 E(p) \quad (3.7.49)$$

and we continue

$$\Leftrightarrow 1 + E(p) = (p - c)f_1 E(p) \quad (3.7.50)$$

$$\Leftrightarrow 1 = (p - c)f_1 E(p) - E(p) \quad (3.7.51)$$

$$\Leftrightarrow 1 = [(p - c)f_1 - 1]E(p) = [(p - c)f_1 - 1]\exp(-f_0 + f_1 \cdot p) \quad (3.7.52)$$

Again, we wish to achieve the form  $ye^y$ , so we put everything else on the LHS of the equation. In particular, we determine the factor of  $p$  in the exponent (which is  $f_1$ ) and replicate the product  $p f_1$  at the base.

$$\exp(f_0) = [p f_1 - 1 - c f_1] \exp(f_1 \cdot p) \quad (3.7.53)$$

Finally, we replicate everything in the base that's not yet in the exponent  $-1 - c f_1$  (by multiplying with the appropriate exponential term)

$$\exp(f_0) \exp(-1 - c f_1) = [p f_1 - 1 - c f_1] \exp(f_1 \cdot p) \exp(-1 - c f_1) \quad (3.7.54)$$

$$\Leftrightarrow \exp(f_0 - 1 - c f_1) = [p f_1 - 1 - c f_1] \exp(p f_1 - 1 - c f_1). \quad (3.7.55)$$

We can now apply Lambert-W, and solve for  $p$

$$W(\exp(f_0 - 1 - c f_1)) = p f_1 - 1 - c f_1 \quad (3.7.56)$$

$$\Leftrightarrow f_1 p = W(\exp(f_0 - 1 - c f_1)) + 1 + c f_1 \quad (3.7.57)$$

to obtain

$$p = \frac{W(\exp(f_0 - 1 - c f_1)) + 1 + c f_1}{f_1}$$

### 3.8 Appendix D: Derivations of Elasticities

In this chapter we derive expressions for determining the elasticity of demand, which we defined as the point elasticity as in (3.3) as  $\epsilon = \frac{\partial Q(p)}{\partial p} \frac{p}{Q(p)}$ .

We make use of the already determined derivatives  $\frac{\partial Q(p)}{\partial p}$  from the previous Chapter 3.7.

- 1.) The linear function  $Q_{lin}(p) = a_0 - a_1 \cdot p$ .

$$\epsilon = \frac{\partial Q(p)}{\partial p} \frac{p}{Q(p)} = \frac{-a_1 p}{a_0 p - a_1} = \frac{a_1 p}{a_1 - a_0 p} \quad (3.8.1)$$

- 2.) The multiplicative function  $Q_{mult}(P) = b_0 \cdot p^{-b_1}$ .

$$\epsilon = \frac{\partial Q(p)}{\partial p} \frac{p}{Q(p)} = \frac{-b_0 b_1 p^{-b_1-1}}{b_0 p^{-b_1}} = -b_1 \quad (3.8.2)$$

- 3.) The exponential function  $Q_{exp}(p) = e^{c_0 - c_1 \cdot p}$ .

$$\epsilon = \frac{\partial Q(p)}{\partial p} \frac{p}{Q(p)} = \frac{-c_1 Q p}{Q} = -c_1 p \quad (3.8.3)$$

- 4.) The semi-logarithmic function  $Q_{semlog}(p) = d_0 - d_1 \cdot \log(p)$ .

$$\epsilon = \frac{\partial Q(p)}{\partial p} \frac{p}{Q(p)} = \frac{-d_1 p}{p Q} = \frac{-d_1}{d_0 - d_1 \cdot \log(p)} = \frac{d_1}{d_1 \cdot \log(p) - d_0} \quad (3.8.4)$$

- 5.) The logistic function  $Q_{log}(p) = \frac{Q_{max}}{1 + e^{-(f_0 - f_1 \cdot p)}} = \frac{Q_{max}}{D(p)}$ .

$$\epsilon = \frac{\partial Q(p)}{\partial p} \frac{p}{Q(p)} = \frac{-Q_{max} E(p)}{D(p)^2} \frac{p D(p)}{Q_{max}} = \frac{-p E(p)}{D(p)} = \frac{-p e^{-(f_0 - f_1 \cdot p)}}{1 + e^{-(f_0 - f_1 \cdot p)}} \quad (3.8.5)$$

$$= \frac{-p}{1 + e^{(f_0 - f_1 \cdot p)}} \quad (3.8.6)$$

## Chapter 4

# Statistical Procedures

In this chapter, we derive the null hypotheses necessary for comparing the allocation procedures from Chapter 2. The article it was based on was accepted for review in the journals ‘Econometrics and Statistics’ and ‘Journal of Applied Statistics’. By the time of submitting this dissertation, it has not been published in a journal. The paper has two authors: Prof. Dr. Harald Hruschka and me. Changes compared to the original submission consist mainly of layout changes, references to the paper from Chapter 2, and deletion of sections necessary only to the journals. We would like to emphasize that some of the proofs and constructions in this chapter are *highly* technical, and that the most complicated concepts are explained in simpler terms in the Appendix Chapter 4.6. The title of the paper as submitted to ‘Econometrics and Statistics’ was:

### **Testing effects of categories of one independent variable in full ANOVA models**

#### ABSTRACT

We construct an exact method to estimate cell means of full ANOVA models defined by inclusion of all main and all interaction effects. To this end we derive estimated cell means based on the comparison of OLS estimated intercepts before and after an appropriate transformation. We further describe how to use cell means to test hypotheses which compare two levels of one independent variable taking all main and interaction effects into account. We conclude with a final comprehensive example. All calculations performed in this article can be checked and customized with three accompanying R files.

Note that the R-files that are mentioned in the paper are not included in this thesis.

## 4.1 Introduction

Multiple linear regression has been the epitome of a research method in the last hundred years in many fields of science. While different circumstances may require other mathematical models, scientists, especially statisticians, still have to master multiple linear regression.

Cohen (1978) specifically explains and demonstrates how to test lower order interactions and powers as independent variables in regression models. However, interpretation of regression models with interaction terms may become quite difficult, especially if models include higher order interaction terms. We deal with multi-way ANOVA, i.e., multiple linear regression with several categorical independent variables (Neter and Wasserman (1974)). There are several different ways of performing ANOVAs. Beside the standard way of manually constructing (sums of squares of) distances to means, one may also consider interpoint distances within and between groups, as demonstrated in Anderson (2001). However, the complexity of such approaches in a multi-factorial-design grows quickly with every additional factor. Interpretation of classical ANOVAs will be highly difficult. In an OLS-setting however, adding a factor just adds columns to the (rather sparse) design matrix. As the equivalence of ANOVAs and regressions with dummy-coded variables is well known (see Jobson (1991), Chapter 5.1.3), we will remain in the setting of the OLS estimator and derive our results with simple linear algebra.

The standard way of comparing effects in an ANOVA is through contrasts. However, Embretson (1996) demonstrated that evaluation contrasts on total scores can lead to spurious interaction effects, with effects showing up when there are none, and vice versa; even in a mere 2x2 factorial design. We therefore use linear algebra to derive appropriate null hypotheses in an OLS setting instead of relying on contrasts.

A full ANOVA model is defined by inclusion of all main and all interaction effects. If every cell, i.e., every combination of values of the independent variables, contributes in a specific way to the dependent variable, a full model may be superior to less complex nested ANOVA models with fewer or even no interactions. If one wants to compare the effects of categories of one selected independent variable on the dependent variable, just looking at their main effects is not appropriate for a model with interactions. A better approach consists in interpreting estimated cell means (see Jobson (1991)). But this approach becomes impractical if the number of cells is high due to many independent variables or many categories. Then we need a summarizing method which is not restricted to main effects, but considers all interaction effects as well. In this article we develop such a method for full ANOVA models. Be warned though, that the presence of more and higher interactions need to be justified by either a large sample size or large expected effect sizes (see Cohen et al (2013), Chapter 9.2).

We start the article proving two Theorems on cell means:

**Theorem.** For a full  $N$ -way dummy-coded ANOVA model, the estimated mean of a cell containing non-reference categories  $(c_1, \dots, c_N)$  is the sum of their interaction coefficient  $\hat{\beta}_{1,c_1;2,c_2;\dots;N,c_N}$  and every coefficient in which only those indices appear, plus the intercept  $\hat{\beta}_0$ .

**Theorem.** For a full  $N$ -way dummy-coded ANOVA model, the estimated mean of a cell containing categories  $(c_1, \dots, c_N)$  is the sum of their non-reference-interaction coefficient  $\hat{\beta}_{(k,c_k)_{c_k \neq 0}}$  and every coefficient in which only those indices appear, plus the intercept  $\hat{\beta}_0$ .

To explain these Theorems we start Section 4.2.1 with a very simple full two-way ANOVA model with only one observation per cell and dichotomous independent variables. We then deal with models of increasing complexity by allowing more observations per cell or multichotomous independent variables. We finish this section by looking at a full  $N$ -way ANOVA model with multichotomous independent variables. In Section 4.2.2 we use these Theorems to derive null hypotheses comparing the effects that different categories of a selected independent variable exert on the dependent variable. In Section 4.3 we present an example which demonstrates the utility of the statements from Sections 4.2.1 and 4.2.2. We give a quick outlook in Section 4.4. Mathematical lemmata used for the derivations are proven in Section 4.5 Appendix A. This article was originally accompanied by four online resources, three of which contain explanations and calculations for the examples in this article, including the example from Section 4.3. The fourth has been turned into Section 4.6 Appendix B for this thesis.

## 4.2 Methodology

### 4.2.1 Interaction Coefficients in full $N$ -way ANOVA models

We begin by explaining the statement and proof of the Theorems for a full two-way ANOVA model with dichotomous independent variables for one observation per cell. We then extend this analysis to designs with multiple observations per cell, first for balanced designs and then unbalanced designs.

We generalize these concepts to a full two-way ANOVA model with multichotomous independent variables and finally a full  $N$ -way ANOVA model with multichotomous independent variables. To simplify we only write variables in the following if we mean independent variables. We call the number of rows (and hence also columns) of a square matrix its *dimension*, and the number of rows of non-square matrices its *length*.

#### Two-way ANOVA with dichotomous variables

**One Observation per Cell** We begin with a simple example, a dataset consisting of two dummy-coded dichotomous variables  $X_1, X_2$  (whose categories we call  $x_1^0, x_1^1$  for  $X_1$  and  $x_2^0, x_2^1$  for  $X_2$ ), and one dependent variable  $y$ . Therefore we have four cells to analyze.

Though it is not strictly necessary for our calculations, we assign real numbers 1, 10, 100 and 1,000 to the dependent variable as shown in Table 4.1.

Table 4.1: Table with observed cell values

$X_1 \backslash X_2$	$x_2^0$	$x_2^1$
$x_1^0$	1	10
$x_1^1$	100	1,000

The values are not important by themselves, we chose them in a way that they differ by at least an order of magnitude to facilitate understanding.

For the simplest situation, consider exactly these four data rows, represented by

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}, y = \begin{pmatrix} 1 \\ 10 \\ 100 \\ 1,000 \end{pmatrix}.$$

We now perform a linear regression with the variables  $X_1, X_2$  and their interaction  $X_1 \cdot X_2$ . Including the intercept, this gives us four coefficients to be estimated, which means we have no residual degrees of freedom. Therefore, the OLS estimator is just the unique solution to the equation system  $y = X\hat{\beta}$ . In particular,  $X$  is invertible and we have

$$\hat{\beta} = (X^T X)^{-1} X^T y = X^{-1} (X^T)^{-1} X^T y = X^{-1} \mathbb{1} y = X^{-1} y$$



with

$$X^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

and we get

$$\hat{\beta} = \begin{pmatrix} 1 \\ 9 \\ 99 \\ 891 \end{pmatrix}.$$

Here the mean of the reference cell  $(x_1^0, x_2^0)$  is represented by the intercept which is just 1, which can be seen from the first entry of  $\hat{\beta}$  and the top left entry in Table 4.1.

If we want a different reference category, we have to change the y-variable to feature the desired cell in the first entry. Say we want to consider  $(x_1^1, x_2^1)$  as the reference cell (currently represented in the fourth entry of  $y$ ). We define the permutation matrix  $R$ , which switches the first and fourth entry of  $y$ :

$$R = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

so we have

$$Ry = R \begin{pmatrix} 1 \\ 10 \\ 100 \\ 1,000 \end{pmatrix} = \begin{pmatrix} 1,000 \\ 10 \\ 100 \\ 1 \end{pmatrix}.$$

Note that  $R = R^T$ , as  $R$  is indeed symmetric.

For our OLS estimator this means

$$\hat{\beta}_{\text{new}} = (X^T X)^{-1} X^T Ry = \begin{pmatrix} 1,000 \\ -990 \\ -900 \\ 891 \end{pmatrix}$$

and its first entry, 1,000, is the value of the new reference cell. Of course, we could have just as well switched the rows of  $X$ , as long as the reference cell value for  $y$  has the same row as the intercept row  $(1, 0, 0, 0)$  of  $X$ . So, we could have instead calculated the OLS estimator for  $RX$  instead of  $X$  to get

$$((RX)^T (RX))^{-1} (RX)^T y = (X^T R^T R X)^{-1} X^T R y = (X^T X)^{-1} X^T R y = \hat{\beta}_{\text{new}}, \quad (4.2.1)$$

as permutation matrices are orthogonal, see Lemma 4.5.1 (4).

*Remark 4.2.1.* Note that we could have also used the matrix

$$R' = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

With  $R'$ , the variables remain the same and both categories change. With  $R$ , the variables swap, *and* they change their categories. Applying  $R'$  to  $X$  is equivalent to redefining the variables  $X'_1 = 1 - X_1$  and  $X'_2 = 1 - X_2$ , whereas applying  $R$  is equivalent to redefining  $\hat{X}_1 = 1 - X_2$  and  $\hat{X}_2 = 1 - X_1$ , thereby switching both the variables and the categories. Independent on which of the methods we choose, the desired cell before the transformation becomes the intercept after the transformation. We choose the former version for a simple reason: we later construct  $R$  for an arbitrarily large number of variables and categories. This task is easy with the first method, as we only switch the first entry of the identity matrix with some other entry, independent on the size of  $X$ , whereas the latter construction gets more complicated when the number of cells increases.

The current goal is to reconstruct the value of 1000 from the original estimate  $\hat{\beta}$ . In other words, using OLS coefficient estimates we want to fill in Table 4.2 with entries just from the estimator.

Table 4.2: Table of cells to be filled in

$X_1 \backslash X_2$	$x_2^0$	$x_2^1$
$x_1^0$		
$x_1^1$		

Obviously, readers might argue that the value 1000, for the bottom right corner of Table 4.2 is just the sum of all the entries in  $\hat{\beta}$ , and they would be right, as:

**Proposition 4.2.2.**

$$(1 \ 1 \ 1 \ 1) \hat{\beta} = (1 \ 0 \ 0 \ 0) \hat{\beta}_{\text{new}}$$

*Proof.* Define the matrix  $V$  as

$$V := X^T R (X^T)^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad (4.2.2)$$

then by Lemma 4.5.2 from Chapter 4.5 Appendix A we know that  $VX^T = X^T R$ ,  $R^T X = XV^T$  and  $V^{-1} = V$ , so we can conclude

$$\begin{aligned} \hat{\beta}_{\text{new}} &= (X^T X)^{-1} X^T R y = (X^T X)^{-1} V X^T y = (V^{-1} X^T X)^{-1} X^T y \\ &= (V X^T X)^{-1} X^T y = (X^T R X)^{-1} X^T y = (X^T R^T X)^{-1} X^T y \\ &= (X^T X V^T)^{-1} X^T y = V^T (X^T X)^{-1} X^T y = V^T \hat{\beta}, \end{aligned} \quad (4.2.3)$$

and indeed

$$(1 \ 1 \ 1 \ 1) \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = (1 \ 0 \ 0 \ 0).$$

□

*Remark 4.2.3.* Because of symmetry this works the other way around as well, as

$$(1 \ 0 \ 0 \ 0) \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = (1 \ 1 \ 1 \ 1)$$

and the other calculations go through without changes. This way of ‘pulling matrices through the OLS estimator’ is our most beneficial tool. In particular, in every generalization step of this method (where necessary), we explain the meaning behind  $R_Y$  and prove that ‘ $R$  can be pulled through the OLS estimator’.

**Balanced Designs** Surely, one could argue that OLS estimates have little meaning if the design matrix is a square matrix. Therefore, in the next steps, we prove the statement of Proposition 4.2.2 under less restrictive conditions. Note that we do not need  $X$  to be invertible in the proof of Proposition 4.2.2, but use the general statement  $VX^T = X^T R$ . For more general design matrices, we require a more general version of the matrices  $X$  and  $R$ .

**Definition 4.2.4.** Let  $X$  be defined as

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

then we call  $X$  *reduced design matrix* and call its rows *usual rows*. All reduced design matrices appearing in this article are constructed according to Definition 4.2.21 in Section 4.2.1. Hitherto we will simply provide the matrices themselves. A *generic design matrix*  $M$  is a matrix that only consists of the usual rows (i.e., only of the rows of the reduced design matrix), each at least once.

To every generic design matrix  $M$  we can define a vector  $a$  whose entries denote the sequence in which the usual rows appear. We call this vector the *manifest vector* of  $M$ . Obviously, this vector is uniquely determined by  $M$  and vice versa. In particular, it holds that  $M_{i,j} = X_{a_i,j}$ . The number of rows in  $M$  (and hence of entries in the manifest vector) is again called its *length*.

If every entry of the manifest vector appears equally often, we call  $M$  *balanced*.

We will focus on balanced designs first, and then discuss unbalanced designs.

**Example 4.2.5.** Of course,  $X$  itself is a generic design matrix, which is balanced. Further, consider the matrices

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, U_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \text{ and } U_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

The manifest vector of  $X$  itself is  $(1, 2, 3, 4)$ , and the manifest vector of  $T$  would be  $(1, 2, 3, 4, 3, 4, 2, 1)$ . The other two matrices have no valid manifest vectors, as  $U_1$  does not include the second usual row, and  $U_2$  has a non-usual sixth row. In particular,  $U_1$  and  $U_2$  are *not* generic design matrices.

Of course, these generic design matrices need to be accompanied by a vector  $y$  of the dependent variable of the same length, where each entry, together with a row from the design matrix represent a joint measurement of the dependent and independent variables. Therefore, the manifest vector also identifies the entries of  $y$  corresponding to these cells.

In order for the statement  $M^T R = V M^T$  to remain true in this situation,  $V$  still has to be a square matrix of dimension 4, and is, in fact, the same matrix as defined above in Equation (4.2.2). If the generic design matrix is of length  $n$ , then  $R$  is a square matrix of dimension  $n$ . It loses the property of being a simple permutation matrix. As long as the design matrix is balanced,  $R$  remains symmetric. We now demonstrate a possible construction, still for the situation of a switch between row 1 and 4, i.e., changing both reference categories.

**Definition 4.2.6.** Let  $M$  be a balanced generic design matrix of length  $n$  and  $a = (a_1, a_2, \dots, a_n)$  its manifest vector. We define the matrix  $R$  by means of the following construction.

Let  $I_j \subset \{1, \dots, n\}$  be the set of positions where  $a_l = j (\Leftrightarrow l \in I_j)$  for  $j \in \{1, 2, 3, 4\}$ . In other words, let  $I_1$  contain precisely those indices  $l$  where  $a_l = 1$ , and so forth.

Step 1:

Let  $R_0$  be the square matrix of dimension  $n$  with columns  $r^1, \dots, r^n$ . For every  $l \in I_1$  the column  $r^l$  has a 1 in every entry from  $I_4$  and a zero in all other positions. For every  $l \in I_4$ ,  $r^l$  has a 1 in every entry from  $I_1$  and a zero in all other positions. Furthermore, for  $k = 2, 3$ , for every  $l \in I_k$ ,  $r^l$  has a 1 in every entry from  $I_k$  and a zero in all other positions.

This makes sure that 1 and 4 get switched, in contrast to 2 and 3, which keep their positions.

By construction,  $R_0$  is symmetric.

Step 2:

Divide the matrix  $R_0$  by  $n/4$ . Call the resulting matrix  $R$ , which is still symmetric.

**Example 4.2.7.** Consider the matrix  $T$  from Example 4.2.5, with the manifest vector  $a = (1, 2, 3, 4, 3, 4, 2, 1)$ . Then we have  $I_1 = \{1, 8\}$ ,  $I_2 = \{2, 7\}$ ,  $I_3 = \{3, 5\}$  and  $I_4 = \{4, 6\}$ .

Step 1:

The columns  $r^1$  and  $r^8$  consist of a 1 in the 4th and 6th position, and zeroes elsewhere.

The columns  $r^2$  and  $r^7$  consist of a 1 in the 2nd and 7th position.

The columns  $r^3$  and  $r^5$  have a 1 in the 3rd and 5th position.

The columns  $r^4$  and  $r^6$  have a 1 in the 1st and 8th position.

The matrix  $R_0$  hence looks like this:

$$R_0 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Step 2:

Divide the matrix by  $n/4 = 8/4 = 2$  to obtain  $R$

$$R = \begin{pmatrix} 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \end{pmatrix}$$

And indeed, it holds that

$$T^T R = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = VT^T,$$

and also

$$Ra = R(1, 2, 3, 4, 3, 4, 2, 1)^T = (4, 2, 3, 1, 3, 1, 2, 4)^T,$$

which could even be interpreted as the manifest vector of the generic design matrix  $(T^T R)^T = R^T T$ .

What remains to be proven is:

**Proposition 4.2.8.** Let  $M$  be a balanced generic design matrix of length  $n$ ,  $a$  its manifest vector and  $y$  another vector of dimension  $n$ . Let  $V$  be as in Equation (4.2.2) and  $R$  be constructed from  $a$  as in Definition 4.2.6. Let  $\hat{\beta} := (M^T M)^{-1} M^T y$  and  $\hat{\beta}_{\text{new}} := (M^T M)^{-1} M^T R^T y$ .

- 1.) Right-multiplication of  $R$  to  $M^T$  replaces the usual column 1 of  $M^T$  with the usual column 4 and vice versa. The other columns remain the same.
- 2.) Left-multiplication of  $V$  to  $M^T$  replaces the usual column 1 of  $M^T$  with the usual column 4 and vice versa. The other columns remain the same.
- 3.)  $M^T R = VM^T$
- 4.)  $R^T M = MV^T$
- 5.)  $\hat{\beta}_1 = \hat{\beta}_{\text{new}_1} + \hat{\beta}_{\text{new}_2} + \hat{\beta}_{\text{new}_4} + \hat{\beta}_{\text{new}_4}$
- 6.)  $\hat{\beta}_{\text{new}_1} = \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$
- 7.) For a full two-way dummy-coded ANOVA model with dichotomous variables, the mean of the cell with non-reference categories is the sum of the OLS coefficients.

*Remark 4.2.9.* Note that by Lemma 4.5.6, all OLS estimators in this article are well-defined, since by our definition of a generic design matrix  $M$ , the matrix  $M^T M$  is invertible.

- Proof.* 1.) The multiplication of  $M^T$  with a column of  $R_0$  belonging to  $I_1$  produces the usual fourth column with a factor of as many times as it appears. The multiplication with a column of  $R$  hence produces the usual fourth column exactly once. The resulting matrix hence has a usual fourth column in every position  $M^T$  had a usual first column. In the same way, every usual fourth column is replaced by the usual first column, and the other column types stay the same.
- 2.) Left-multiplication of  $V$  to  $X^T$  replaces the usual column 1 of  $X^T$  with the usual column 4 and vice versa. The other columns remain the same. Hence, by the definition of matrix multiplication, it does so with every matrix containing the same columns as  $X^T$ , and therefore with  $M^T$ .
- 3.) Follows immediately from (1) and (2).
- 4.) Like in Lemma 4.5.2

$$R^T M = (M^T (R^T)^T)^T = (M^T R)^T \stackrel{3.}{=} (VM^T)^T = MV^T.$$

- 5.) Same as in Equation (4.2.3),

$$\begin{aligned} V^T \hat{\beta} &= V^T (M^T M)^{-1} M^T y = (M^T M (V^T)^{-1})^{-1} M^T y = (M^T M V^T)^{-1} M^T y \\ &= (M^T R^T M)^{-1} M^T y = (M^T R M)^{-1} M^T y = (V M^T M)^{-1} M^T y \\ &= (M^T M)^{-1} (V)^{-1} M^T y = (M^T M)^{-1} V M^T y \\ &= (M^T M)^{-1} M^T R y = \hat{\beta}_{\text{new}} \end{aligned}$$

and hence,

$$\begin{aligned} \hat{\beta}_1 &= (1, 0, 0, 0) \hat{\beta} = (1, 1, 1, 1) V^T \hat{\beta} = (1, 1, 1, 1) \hat{\beta}_{\text{new}} \\ &= \hat{\beta}_{\text{new}_1} + \hat{\beta}_{\text{new}_2} + \hat{\beta}_{\text{new}_4} + \hat{\beta}_{\text{new}_4} \end{aligned}$$

6.) Again,

$$\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 = (1, 1, 1, 1)\hat{\beta} = (1, 0, 0, 0)V^T\hat{\beta} = (1, 0, 0, 0)\hat{\beta}_{\text{new}} = \hat{\beta}_{\text{new}_1}$$

7.) The mean of the cell containing non-reference categories is equal to the intercept if 1 and 4 are swapped. This is precisely what has been proven in (1) - (6).  $\square$

*Remark 4.2.10.* If instead one wishes to know the mean of a different cell, all that is needed is to switch different entries. Starting with the transposition of indices  $i, j$  one wishes to consider and construct the ‘reduced’ 4 by 4 transition matrix  $R_{\text{red}}$ . Define  $V$  as  $X^T R_{\text{red}}(X^T)^{-1}$  where  $X$  is the reduced design matrix.

Finally, construct the actual matrix  $R$  in analogy to Definition 4.2.6 in order to prove equality. For the effects vector, simply multiply  $(1, 0, 0, 0)V^T\hat{\beta}$ .

For example, in order to find the effect of the reference category in the first variable and the non-reference-category in the second variable, one needs to switch indices 1 and 3. In this situation,  $R_{\text{red}}$  and  $V$  have the following shapes:

$$R_{\text{red}} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, V = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Therefore, since  $(1, 0, 0, 0)V^T = (1, 0, 1, 0)$ , the effect can be reconstructed from the original OLS coefficients by adding  $\hat{\beta}_1$  and  $\hat{\beta}_3$ .

*Remark 4.2.11.* Again, note that we could have instead chosen the matrix

$$R'_{\text{red}} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

We however always choose the construction that switches the first entry with one other entry of an identity matrix and leave everything else the way it is.

The statement of Proposition 4.2.8 may seem to be generally known, obvious even. But the existence of the matrix  $R$  that can be pulled through the OLS estimator is not obvious.

We can now fill Table 4.2 with elements of  $\hat{\beta}$  only, yielding Table 4.3

Table 4.3: Filled in Table

$X_1 \setminus X_2$	$x_2^0$	$x_2^1$
$x_1^0$	$\hat{\beta}_1$	$\hat{\beta}_1 + \hat{\beta}_3$
$x_1^1$	$\hat{\beta}_1 + \hat{\beta}_2$	$\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$

**Unbalanced Designs** We now consider the general unbalanced situation. The calculations are a bit more involved. We start with a generic design matrix  $M$  together with its reduced version  $X$  and construct the same matrices  $R_{\text{red}}$  and  $V$ . However,  $R$  is not symmetric anymore and the deviation from being symmetric is measured by a diagonal square matrix. In particular we see:

**Proposition 4.2.12.** *For every generic design matrix  $M$  with manifest vector  $a$  there is a matrix  $R$  that has the following properties:*

- 1.) *The product  $R^T a$  switches the entries of  $a$  labeled 1 and 4, and leaves entries labeled 2 and 3 fixed.*
- 2.) *The product  $R^T M$  switches the usual first and fourth rows, and leaves the usual second and third rows fixed.*
- 3.) *For  $V := X^T R_{\text{red}} (X^T)^{-1}$  it holds that  $VM^T = M^T R$ .*
- 4.) *There is a diagonal matrix  $D$  such that  $R^T D = R$ .*
- 5.) *There is a matrix  $V_2$  such that  $V_2 M^T = M^T D^{-1}$ .*
- 6.) *The product  $R^T y$  replaces the entries of a dependent variable  $y$  belonging to the first usual row with the mean of the entries belonging to the fourth usual row and vice versa. It further replaces the entries belonging to the second and third usual row with their respective mean.*

Note that replacing values of  $y$  that have equal rows in  $M$  with their mean does not change the outcome of the OLS estimator, since

**Lemma 4.2.13.** *Let  $M$  be a matrix (of full rank) with two equal rows  $i_1$  and  $i_2$ , in other words  $M_{i_1 j} = M_{i_2 j} := \tilde{M}_j$  for all  $j$  and  $y$  be a vector of the same length. Then  $M^T y$  (and hence  $(M^T M)^{-1} M^T y$ ) depends on the sum  $y_{i_1} + y_{i_2}$  but not their individual values.*

*Proof.*

$$\begin{aligned} (M^T y)_j &= \sum_k (M^T)_{jk} y_k = \sum_k M_{kj} y_k = M_{i_1 j} y_{i_1} + M_{i_2 j} y_{i_2} + \sum_{k \neq i_1, i_2} M_{ki} y_k \\ &= \tilde{M}_j y_{i_1} + \tilde{M}_j y_{i_2} + \sum_{k \neq i_1, i_2} M_{ki} y_k = (y_{i_1} + y_{i_2}) \tilde{M}_j + \sum_{k \neq i_1, i_2} M_{ki} y_k \end{aligned}$$

□

**Remark 4.2.14.** Lemma 4.2.13 has an obvious generalization for  $l > 2$  equal rows and the same proof.

After constructing  $R$  (and  $R^T$ ) below in 4.2.6, we prove Proposition 4.2.12, by proving 6.) in Lemma 4.5.4 in the Appendix. It is easily seen, that 1.) and 2.) directly follow from 6.). The proof of 4.) can be found in Lemma 4.5.4 as well, as both proofs are quite technical. Finally, 3.) and 5.) follow with the same arguments as in Proposition 4.2.8.



**Definition 4.2.15.** Let  $M$  be a generic design matrix of length  $n$  and  $a = (a_1, a_2, \dots, a_n)$  its manifest vector. Let  $X$  be the corresponding reduced matrix, which is square of dimension  $m < n$  (for now, still  $m = 4$ ). Without loss of generality we may assume that the first  $n$  rows of  $M$  are identical to  $X$ . Otherwise we define a permutation matrix  $P$  such that the manifest vector of  $PM$ , which is again a generic design matrix, does so. This is always possible, because generic design matrices contain each row of  $X$  at least once. Then we use Equation (4.2.1) and continue calculating with  $M' = PM$ .

Let  $i_1, i_2$  be the indices we intend to swap, then we can disjointly decompose the set  $I := \{1, \dots, n\}$  into  $m$  sets  $I_j$ .

Let  $I_j \subset I$  be the set of positions where  $a_l = j \Leftrightarrow l \in I_j$  for  $j \in \{1, 2, 3, 4\}$ . In other words, let  $I_1$  contain precisely those indices  $l$  where  $a_l = 1$ , and so forth.

Step 1:

Let  $R_0$  be the  $n$  by  $n$  matrix with columns  $r^1, \dots, r^n$ . For every  $l \in I_{i_1}$  the column  $r^l$  has a 1 in every entry from  $I_{i_2}$ . For every  $l \in I_{i_2}$ ,  $r^l$  has a 1 in every entry from  $I_{i_1}$ . Furthermore, for  $k \in I \setminus \{i_1, i_2\}$ , for every  $l \in I_k$ ,  $r^l$  has a 1 in every entry from  $I_k$ .

This makes sure that only  $i_1$  and  $i_2$  get switched.

Note that  $R_0$  is still symmetric.

Step 2:

Divide each column of  $R_0$  by the sum of its entries. Call the resulting non-symmetric matrix  $R$ . We can also divide every row of  $R_0$  by the sum of its entries. Since  $R_0$  is symmetric with non-zero row-sums, this gives us  $R^T$ . In particular, by Lemma 4.5.4 there is a diagonal matrix  $D$  such that  $D^{-1}R^T = R = R^TD$ .

As all the information of  $X$  is contained in the first rows of  $M$ , we define  $D_{\text{top}}$  to be the top-left square submatrix of  $D$  of dimension  $m$ . We then define  $V_2 := X^T D_{\text{top}} X^{T-1}$  and by the same arguments as in Proposition 4.2.8 (1.) and (2.), we have  $V_2 M^T = M^T D^{-1}$ .

**Theorem 4.2.16.**  $V^T \hat{\beta} = (M^T M)^{-1} M^T R^T y$

*Proof.*

$$\begin{aligned}
 V^T \hat{\beta} &= V^T (M^T M)^{-1} M^T y = (M^T M (V^T)^{-1})^{-1} M^T y \\
 &= (M^T M V^T)^{-1} M^T y = (M^T R^T M)^{-1} M^T y \\
 &= (M^T R D^{-1} M)^{-1} M^T y = (V M^T D^{-1} M)^{-1} M^T y \\
 &= (V V_2 M^T M)^{-1} M^T y = (M^T M)^{-1} V_2^{-1} V^{-1} M^T y \\
 &= (M^T M)^{-1} V_2^{-1} V M^T y = (M^T M)^{-1} V_2^{-1} M^T R y \\
 &= (M^T M)^{-1} V_2^{-1} M^T D^{-1} R^T y = (M^T M)^{-1} V_2^{-1} V_2 M^T R^T y \\
 &= (M^T M)^{-1} M^T R^T y
 \end{aligned}$$

□

**Remark 4.2.17.** Note that  $R^T$  is nothing but a notation. We might have just as well defined  $S := R^T$  to get a statement not including a transpose sign. The statement itself is powerful: **even in unbalanced designs, the same matrix  $V^T$  as in balanced designs leads to the linear combination necessary for determining the mean of the desired cell. In particular:**

$$\hat{\beta}_{\text{new}_1} = \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4$$

**Example 4.2.18.** Consider a matrix  $T$  with the manifest vector  $a = (1, 3, 2, 3, 3, 4, 2, 1)^T$ . Define the matrix  $P$  as

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

which leads to  $Pa = (1, 2, 3, 4, 3, 3, 2, 1)^T$ .

Then we have  $I_1 = \{1, 8\}$ ,  $I_2 = \{2, 7\}$ ,  $I_3 = \{3, 5, 6\}$  and  $I_4 = \{4\}$ .

Step 1:

The columns  $r^1$  and  $r^8$  consist of a 1 in the 4th position, and zeroes elsewhere.

The columns  $r^2$  and  $r^7$  consist of a 1 in the 2nd and 7th position.

The columns  $r^3$ ,  $r^5$  and  $r^6$  have a 1 in the 3rd, 5th and 6th position.

The column  $r^4$  has a 1 in the 1st and 8th position.

The matrix  $R_0$  hence looks like this:

$$R_0 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Step 2:

Divide each column by the sum of its entries to yield:

$$R = \begin{pmatrix} 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \end{pmatrix}$$

And indeed, it holds that

$$(PT)^T R = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = V(PT)^T,$$

and also

$$Ra = R(1, 2, 3, 4, 3, 3, 2, 1)^T = (4, 2, 3, 1, 3, 3, 2, 4)^T,$$

which could again be interpreted as the manifest vector of the generic design matrix  $(T^T R)^T = R^T T$ . With the same argument as in Equation (4.2.1), we see that multiplying  $PT$  is equivalent to multiplying  $Py$ , so we define  $T' := PT$  and continue our calculation. The diagonal matrix  $D$  is

$$D := \text{diag}(2, 1, 1, 1/2, 1, 1, 1, 2),$$

and we have  $R = R^T D$ . We define the  $D_{\text{top}}$  to be the top left corner submatrix of  $D$  of dimension 4:

$$D_{\text{top}} = \text{diag}(2, 1, 1, 1/2)$$

to construct

$$V_2 := X^T D_{\text{top}} X^{T^{-1}} = \begin{pmatrix} 2 & -1 & -1 & 0.5 \\ 0 & 1 & 0 & -0.5 \\ 0 & 0 & 1 & -0.5 \\ 0 & 0 & 0 & 0.5 \end{pmatrix}$$

and indeed we have  $V_2 T'^T = T'^T D$ .

### Two-way ANOVA with multichotomous variables

We now increase the number of categories for each variable, and later, increase the number of variables. First of all, we need a change in notation. We still call our variables  $X_1$  and  $X_2$ . For the remainder of this section,  $X_1$  has  $n_1 > 1$  categories and  $X_2$  has  $n_2 > 1$  categories. The reference categories are  $x_1^0$  and  $x_2^0$ , respectively. The other categories are  $x_1^1, \dots, x_1^{n_1-1}$  and  $x_2^1, \dots, x_2^{n_2-1}$ . We proceed in this manner mainly because we want the reference categories to carry the index 0, and also because we want the estimated coefficients to begin with index 1.

Again, we perform an OLS estimation with both variables and *all interactions*. The number of coefficients to be estimated is hence 1 [intercept] +  $n_1 - 1$  [main effects  $X_1$ ] +  $n_2 - 1$  [main effects  $X_2$ ] +  $(n_1 - 1)(n_2 - 1)$  [interaction terms], which is

$$1 + n_1 - 1 + n_2 - 1 + n_1 n_2 - n_1 - n_2 + 1 = n_1 n_2. \quad (4.2.4)$$

This is precisely the total number of cells, which is no coincidence as shown at the end of this article in Lemma 4.5.5 in Chapter 4.5 Appendix A.

In the coefficient vector,  $\hat{\beta}_0$  denotes the intercept,  $\hat{\beta}_{1,l}$  the coefficient belonging to category  $x_1^l$  and  $\hat{\beta}_{2,k}$  the coefficient belonging to  $x_2^k$  for all  $l \in \{1, \dots, n_1\}$  and  $k \in \{1, \dots, n_2\}$ .

The interaction term of the former two variables carries the coefficient  $\hat{\beta}_{1,l;2,k}$ . We emphasize at this point that we have deliberately chosen the order in which coefficients appear. While order is not crucial here, it turns out to be important if we have more than two variables.

We wish to fill Table 4.4

only with entries from  $\hat{\beta}$ , just like we did with Table 4.2.

We know that the top left corner is  $\hat{\beta}_0$  and there is an obvious guess for a generalization

Table 4.4: Table of cells to be filled in

$X_1 \backslash X_2$	$x_2^0$	$x_2^1$	...	$x_2^{n_2-1}$
$x_1^0$				
$x_1^1$				
...				
$x_1^{n_1-1}$				

of the previous section which we can easily prove by the same methods as in Section 4.2.1.

For brevity, we prove, that switching from reference cell  $(x_1^0, x_2^0)$  to the cell  $(x_1^l, x_2^k)$  for  $k, l \geq 1$  results in the sum  $\hat{\beta}_0 + \hat{\beta}_{1,l} + \hat{\beta}_{2,k} + \hat{\beta}_{1,l;2,k}$ . In other words, the cell in row  $x_1^l$  and column  $x_2^k$  in Table 4.4 corresponds to this sum.

Since we impose an order on the variables, we also impose an order on the (reduced) design matrix. We define as reduced design matrix the following matrix (wherein  $\iota$  denotes the intercept column)

$$X = \begin{pmatrix} \iota & x_1^1 & x_2^1 & \dots & x_1^{n_1-1} & x_2^1 & \dots & x_2^{n_2-1} & x_1^1 x_2^1 & \dots & x_1^{n_1-1} x_2^{n_2-1} \\ (x_1^0, x_2^0) \rightarrow & 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ (x_1^1, x_2^0) \rightarrow & 1 & 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ (x_2^1, x_2^0) \rightarrow & 1 & 0 & 1 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \dots & & & & & & & & & & & \\ (x_1^{n_1-1}, x_2^0) \rightarrow & 1 & 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ (x_1^0, x_2^1) \rightarrow & 1 & 0 & 0 & \dots & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ (x_1^1, x_2^1) \rightarrow & 1 & 1 & 0 & \dots & 0 & 1 & \dots & 0 & 1 & \dots & 0 \\ \dots & & & & & & & & & & & \\ (x_1^{n_1-1}, x_2^{n_2-1}) \rightarrow & 1 & 0 & 0 & \dots & 0 & 0 & \dots & 1 & 0 & \dots & 1 \end{pmatrix}$$

(for more details, see the general construction in Definition 4.2.21). As shown in Equation (4.2.4), this is a square matrix of dimension  $n_1 n_2$ .

We wish to change variable  $X_1$  from the reference category to  $x_1^l$  and variable  $X_2$  from the reference category to  $x_2^k$ . In the reduced situation, this is accomplished by defining the square matrix  $R_{\text{red}}$  of dimension  $n_1 n_2$  which switches the entries 1 and  $1 + n_1 k + l$  and looks as follows:

$$R_{\text{red}} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & & & & & & & & \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 1 + n_1 k + l \rightarrow & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & & & & & & & & \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Again, we define  $V = X^T R_{\text{red}} (X^T)^{-1}$ , which is possible since  $X$  is always invertible

(see Lemma 4.5.3) as a reduced design matrix. Again, we have

$$XV^T = X(X^T R_{\text{red}}(X^T)^{-1})^T = XX^{-1}R_{\text{red}}^T(X^T)^T = \mathbb{1}R_{\text{red}}^T X = R_{\text{red}}X,$$

which is the operation on  $X$  that switches the aforementioned two rows, while keeping the other rows constant. Furthermore

$$V^2 = X^T R_{\text{red}}(X^T)^{-1} X^T R_{\text{red}}(X^T)^{-1} = X^T R_{\text{red}} R_{\text{red}}(X^T)^{-1} = X^T (X^T)^{-1} = \mathbb{1}.$$

In particular, row  $1 + n_1 k + l$  is the row of  $(x_1^l, x_2^k)$ , meaning its product with  $\hat{\beta}$  gives  $\hat{\beta}_0 + \hat{\beta}_{1,l} + \hat{\beta}_{2,k} + \hat{\beta}_{1,l;2,k}$  and  $(1, 0, \dots, 0)V^T \hat{\beta}$  is equal to this row.

All that remains to be shown, is that for a generic design matrix  $M$ , a matrix  $R$  can be constructed such that  $R^T M = M V^T$ .

But the construction in Definition 4.2.15 is universal, in the sense that it does not depend on the number of categories in each variable:

Given a generic design matrix  $M$  of length  $m \geq n_1 n_2$  the manifest vector  $a$  now has tuples  $(x_1^l, x_2^k)$  as entries corresponding to all cells. These entries contain the elements of the row indexing to the left of the ‘ $\rightarrow$ -arrows’ in the reduced design matrix  $X$ . We can again decompose the manifest vector into the disjoint union of sets  $I_{c,d}$  corresponding to the rows from the reduced design matrix where they appear.

In step 1, we construct a square matrix  $R_0$  of dimension  $m$  by finding its columns which are the entries of  $a$  belonging to  $I_{0,0}$  and placing a 1 in every row of this vector that has an index belonging to  $I_{l,k}$ . We place a 1 in every column belonging to  $I_{l,k}$  of  $R_0$  at the row-indices belonging to  $I_{0,0}$ . We proceed by filling in the rest corresponding to their own columns of rows.

In step 2, we divide each column by the sum of its entries to obtain  $R$ .

Again, by construction, we have  $M V^T = R^T M$ . Also, from the construction of  $R$ , we get a diagonal matrix  $D$  such that  $R = R^T D$ . We can once more define the top left diagonal submatrix  $D_{\text{top}}$  of  $D$  and define  $V_2 := X^T D_{\text{top}} X^T^{-1}$  which gives  $V_2 M^T = M^T R^T$ . We repeat the steps of Theorem 4.2.16 and we are done.

**Proposition 4.2.19.** *In a dummy-coded full 2-way ANOVA with multichotomous variables, the mean of the cell with non-reference-categories  $(x_1^k, x_2^l)$  is the sum of their interaction coefficient  $\hat{\beta}_{1,k;2,l}$  and every coefficient in which at most these indices appear ( $\hat{\beta}_{1,k} + \hat{\beta}_{2,l}$ ), plus the intercept  $\hat{\beta}_0$ .*

*Proof.* See above. □

**Remark 4.2.20.** We can now fill in the Table 4.5 as required.

Also, the construction of  $R_{\text{red}}, V, R_0, R, D$  and  $V_2$  did not actually rely on the fact that both variables assume non-reference values. The same calculations hold when one of the variables remains in the reference category, with the results seen in the Table 4.5.

Table 4.5: Complete table

$X_1 \setminus X_2$	$x_2^0$	$x_2^1$	...	$x_2^{n_2-1}$
$x_1^0$	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_{2,1}$	...	$\hat{\beta}_0 + \hat{\beta}_{2,n_2-1}$
$x_1^1$	$\hat{\beta}_0 + \hat{\beta}_{1,1}$	$\hat{\beta}_0 + \hat{\beta}_{1,1} + \hat{\beta}_{2,1} + \hat{\beta}_{1,1;2,1}$	...	$\hat{\beta}_0 + \hat{\beta}_{1,1} + \hat{\beta}_{2,n_2-1} + \hat{\beta}_{1,1;2,n_2-1}$
...	...	...	...	...
$x_1^{n_1-1}$	$\hat{\beta}_0 + \hat{\beta}_{1,n_1-1}$	$\hat{\beta}_0 + \hat{\beta}_{1,n_1-1} + \hat{\beta}_{2,1} + \hat{\beta}_{1,n_1-1;2,1}$	...	$\hat{\beta}_0 + \hat{\beta}_{1,n_1-1} + \hat{\beta}_{2,n_2-1} + \hat{\beta}_{1,n_1-1;2,n_2-1}$

### ***N*-way ANOVA with multichotomous variables**

Up to now, the results we have produced are familiar in some way or another. We now tackle the general situation of more than two variables.

Extending the notation from the previous section, we now have  $N$  dummy-coded variables  $X_1, \dots, X_N$  with  $n_1, \dots, n_N$  categories. Single categories are denoted by  $x_p^q$  meaning that the variable  $X_p$  has the category  $q$ , where categories are counted from 0 to  $n_p - 1$ .

The reference case is therefore represented by the vector  $(x_1^0, \dots, x_N^0)$ . We now define the reduced design matrix  $X$  for the most general case:

**Definition 4.2.21.** Given  $N$  dummy variables  $X_1, \dots, X_N$  with  $n_1, \dots, n_N$  categories respectively, we construct the reduced design matrix  $X$  of dimension  $\tilde{N} = \prod_{i=1}^N n_i$  as follows. We first explain how to index the rows and columns. We index the rows according to the cells. Therefore the first row is indexed with the cell  $(x_1^0, x_2^0, \dots, x_N^0)$ , and the  $k + 1$ -th row with cell  $(x_1^{k_1}, x_2^{k_2}, \dots, x_N^{k_N})$  where  $k_1 := k \bmod n_1$ ,  $k_2 = [(k - (k_1))/n_1] \bmod n_2$ ,  $k_3 = ((k - (k_1))/n_1 - k_2)/n_2 \bmod n_3$  and so forth.

Sometimes we simply write  $(k_1, \dots, k_N)$  instead of  $(x_1^{k_1}, x_2^{k_2}, \dots, x_N^{k_N})$ .

We index the columns according to their position in the OLS estimator. The first column represents the intercept  $\hat{\beta}_0$ , the next  $(n_1 - 1) + \dots + (n_N - 1)$  columns represent the non-reference main effects of the variables  $X_1, \dots, X_N$  where  $\hat{\beta}_{i,a}$  is the effect of variable  $X_i$  being in category  $a$ . These columns are indexed from left to right, i.e. starting with categories, then the variables, i.e.  $\hat{\beta}_{1,n_1-2}$  before  $\hat{\beta}_{1,n_1-1}$  before  $\hat{\beta}_{2,1}$ .

The next  $\sum_{I \subset \{1, \dots, N\}, \#I=2} \prod_{i \in I} (n_i - 1)$  columns are the pairwise interactions of two variables  $X_i X_j$  represented by  $\hat{\beta}_{i,a;j,b}$  where we count up the categories  $b$  and  $a$ , then the variables  $j$  and  $i$ . In other words  $\hat{\beta}_{1,n_1-2;2,n_2-2}$  precedes  $\hat{\beta}_{1,n_1-2;2,n_2-1}$  precedes  $\hat{\beta}_{1,n_1-1;2,1}$  and  $\hat{\beta}_{1,n_1-1;2,n_2-1}$  precedes  $\hat{\beta}_{1,1;3,1}$ . Continue with higher interaction terms, by counting categories before variables, each from right to left.

Now define the entry of  $X$  in row  $k + 1 = (k_1, \dots, k_N)$  and column  $j = \beta_{i_1, a_1; \dots, i_l, a_l}$  to be 1 iff for all  $m = 1, \dots, l$  we have that  $k_{i_m} = a_m$ . Define every other entry to be zero. In particular, for the intercept  $\hat{\beta}_0$ , this is a condition over the empty set, and hence always true. Therefore the first column consists entirely of ones.

We perform a regression of a dependent variable  $y$  against all variables and *all* interactions. In other words the regression model not only includes pairwise, but also triple, quadruple, . . . , up to  $N$ -fold interactions. First of all, we determine the number of coefficients:

$$\begin{aligned}
& 1 \text{ [intercept]} \\
& + (n_1 - 1) + (n_2 - 1) + \dots + (n_N - 1) \text{ [main effects]} \\
& + (n_1 - 1)(n_2 - 1) + (n_1 - 1)(n_3 - 1) + \dots + (n_1 - 1)(n_N - 1) + \dots + (n_{N-1} - 1)(n_N - 1) \text{ [pairwise interactions]} \\
& + (n_1 - 1)(n_2 - 1)(n_3 - 1) + \dots + (n_{N-2} - 1)(n_{N-1} - 1)(n_N - 1) \text{ [triple interactions]} \\
& + \dots \text{ [quadruple to } (N-1)\text{-fold interactions]} \\
& (n_1 - 1)(n_2 - 1)(n_3 - 1) \dots (n_{N-2} - 1)(n_{N-1} - 1)(n_N - 1) \text{ [N-fold interactions]} \\
& = 1 + \sum_{i=1}^N (n_i - 1) + \sum_{i>j\geq 1}^N (n_i - 1)(n_j - 1) + \dots + (n_1 - 1) \dots (n_N - 1) \\
& = \sum_{I \subseteq \{1, \dots, N\}} \prod_{i \in I} (n_i - 1) \stackrel{!}{=} \prod_{i=1}^N n_i =: \tilde{N}
\end{aligned}$$

The proof of the last equality can be found in Lemma 4.5.5.

The reduced design matrix  $X$  is therefore a square matrix of dimension  $\tilde{N}$ , and furthermore invertible (see 4.5.3 for a proof).

We wish to reconstruct the effect of the cell  $(x_1^{c_1}, x_2^{c_2}, \dots, x_N^{c_N})$ . We assume that none of the  $c_i$  are zero, hence we switch to non-reference categories for each variable. Again, the other cases are obvious generalizations, since they require, if anything, less work.

We consider a generic design matrix  $M$ , which now has  $\tilde{N}$  columns and  $m \geq \tilde{N}$  rows. It is accompanied by a manifest vector  $a$  whose entries are  $N$ -tuples containing the categories each row of  $M$  represents.

To switch from the reference case to  $(c_1, \dots, c_N)$ , we need to construct the matrix  $R_{\text{red}}$  that switches the entries 1 and

$$1 + n_1 n_2 \dots n_{N-2} n_{N-1} c_N + n_1 n_2 \dots n_{N-2} c_{N-1} + \dots + n_1 n_2 c_3 + n_1 c_2 + c_1. \quad (4.2.5)$$

(As can be seen in the indexing of rows in Definition 4.2.21, this exactly gives the row of  $(c_1, \dots, c_N)$ ). We define  $V$  as the square matrix of dimension  $\tilde{N}$  which we get by multiplying  $X^T R_{\text{red}} (X^T)^{-1}$ . Note that this is enough to show that the mean of the desired cell is  $(1, 0, 0, 0, \dots, 0) V^T \hat{\beta}$ , where  $(1, 0, 0, 0, \dots, 0) V^T$  has a 1 in every position corresponding to  $\hat{\beta}_{k_1, c_{k_1}; \dots; k_r, c_r}$  for  $k_1 < \dots < k_r$ , so every coefficient in which at least one of the categories  $(c_1, \dots, c_N)$  appear (and the intercept). In other words:

**Theorem 4.2.22.** *For a full  $N$ -way dummy-coded ANOVA model, the estimated mean of a cell containing non-reference categories  $(c_1, \dots, c_N)$  is the sum of their interaction coefficient  $\hat{\beta}_{1, c_1; 2, c_2; \dots, N, c_N}$  and every coefficient in which only those indices appear, plus the intercept  $\hat{\beta}_0$ .*

**Theorem 4.2.23.** *For a full  $N$ -way dummy-coded ANOVA model, the estimated mean of a cell containing categories  $(c_1, \dots, c_N)$  is the sum of their non-reference-interaction coefficient  $\hat{\beta}_{(k, c_k)_{c_k \neq 0}}$  and every coefficient in which only those indices appear, plus the intercept  $\hat{\beta}_0$ .*

We prove the first theorem, the second is an exercise left to the reader. All that needs to change is the index which is to be swapped against 1, which can be determined by Equation (4.2.5).

*Proof.* All that remains is to construct the matrix  $R$  such that  $VM^T = M^T R$  and repeat the usual calculation. Again, we can construct the matrices  $R_0$  and  $R$  with the same methods as in Definition 4.2.15. Decompose the manifest vector  $a$  into subsets  $I_{i_1, i_2, \dots, i_N}$  each corresponding to all the possible entries in  $a$ . This time, in the first step identify the entries which belong to  $I_{0,0,\dots,0}$  and  $I_{c_1, \dots, c_N}$ , in the  $I_{0,0,\dots,0}$ -columns place a 1 in the  $I_{c_1, \dots, c_N}$ -rows and vice versa, finally for all other indices  $I_v$  place a 1 in the  $I_v$  rows of the  $I_v$  columns. Call this matrix  $R_0$ , divide the columns by the sum of their entries and call this matrix  $R$ . By construction,  $VM^T = M^T R$ , and we continue with the usual constructions of  $D$  and  $V_2$  to get  $V_2 M^T = M^T D^{-1}$ . The rest follows just like in Theorem 4.2.16.  $\square$

We now present an example with three variables with two, three and two categories, respectively.

**Example 4.2.24.** Consider the three dummy-coded variables  $X_1, X_2, X_3$  with categories  $x_1^0, x_1^1, x_2^0, x_2^1, x_3^0$  and  $x_3^1$ . The reduced design matrix  $X$  is a square matrix of dimension 12 and has the form:

$$\text{category} \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_{1,1} & \hat{\beta}_{2,1} & \hat{\beta}_{2,2} & \hat{\beta}_{3,1} & \hat{\beta}_{1,1,2,1} & \hat{\beta}_{1,1,2,2} & \hat{\beta}_{1,1,3,1} & \hat{\beta}_{2,2,3,1} & \hat{\beta}_{2,2,3,1} & \hat{\beta}_{1,1,2,1,3,1} & \hat{\beta}_{1,1,2,2,3,1} \end{pmatrix}$$

If we wish to obtain the cell mean of  $(x_1^1, x_2^1, x_3^1)$ , we need to construct the  $R_{\text{red}}$  matrix that swaps the rows 1 and

$$1 + n_1 n_2 c_3 + n_1 c_2 + c_1 = 1 + 2 \cdot 3 \cdot 1 + 2 \cdot 1 + 1 = 1 + 6 + 2 + 1 = 10$$

which looks like this:

[illegible]



We define  $V = X^T R_{\text{red}}(X^T)^{-1}$  which is

$$V = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -2 & -1 \\ 1 & -1 & 0 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -2 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & -1 & 0 & 1 & 1 & 1 & 1 & 1 & -2 & -1 \\ 1 & -1 & -1 & -1 & -1 & 2 & 1 & 1 & 1 & 1 & -2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 2 & 1 & 1 & -2 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 2 & 1 & -2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and finally we can confirm

$$(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)V^T = (1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0),$$

multiplied with  $\hat{\beta}$  gives us

$$\hat{\beta}_0 + \hat{\beta}_{1,1} + \hat{\beta}_{2,1} + \hat{\beta}_{3,1} + \hat{\beta}_{1,1;2,1} + \hat{\beta}_{1,1;3,1} + \hat{\beta}_{2,1;3,1} + \hat{\beta}_{1,1;2,1;3,1}, \quad (4.2.6)$$

confirming the statement of Theorem 4.2.22 for this example.

For reasons that we explain in the next section, we also consider the switch from the reference cell to  $(x_1^0, x_2^1, x_3^1)$ , which is achieved by switching the first entry and entry number

$$1 + n_1 n_2 c_3 + n_1 c_2 + c_1 = 1 + 2 \cdot 3 \cdot 1 + 2 \cdot 1 + 0 = 1 + 6 + 2 = 9.$$

The result is

$$\hat{\beta}_0 + \hat{\beta}_{2,1} + \hat{\beta}_{3,1} + \hat{\beta}_{2,1;3,1}, \quad (4.2.7)$$

confirming Theorem 4.2.23. Note that the coefficients from Equation (4.2.7) also appear in Equation (4.2.6), and are precisely those which do not contain  $(1, 1)$  in their index.

### 4.2.2 Derivation of null hypotheses

Based on Theorems 4.2.22 and 4.2.23, we now derive null hypotheses for a specific question. Let us compare the performance of several methods given certain circumstances, for example the times different cars drive to complete different tracks under different weather conditions, or the solution quality of different algorithms depending on the ‘difficulty’ of the input. Furthermore, assume that a full ANOVA model is preferred to nested smaller models on the basis of F-tests referring to subsets of coefficients Greene (1993). Our goal is by means of the two theorems to bring order to the ‘wall of coefficients’  $\hat{\beta}$ .

Without loss of generality, we assume that the categories of first independent variable are different methods. We denote this variable by  $P_1$  instead of  $X_1$  and call it the *primary variable*. We want to compare the reference category ( $p_1^0$ ) of the primary variable, which we call our *own category*, to the first non-reference-category ( $p_1^1$ ), which we call *competitor category*. This comparison takes place across all cells. All the other independent variables are called *secondary variables* and are indexed by  $S_2, \dots, S_N$  with categories ( $s_i^j$ ). We call the dependent variable  $y$  the *performance variable*.

### Null Hypotheses on Two Cell Means

We start by comparing means of two cells containing secondary variables ( $s_2^{c_2}, s_3^{c_3}, \dots, s_N^{c_N}$ ) in non-reference-categories, i.e.,  $c_k \neq 0 \forall_k$ . From Theorem 4.2.22, we know that the mean of the cell with these values of the secondary variables and the competitor category for the primary variables is given by the sum of all coefficients which hold only these indices ( $c_1 = 1, c_2, c_3, \dots, c_N$ ). By Theorem 4.2.23 the mean of the cell with the same values for the secondary variables and the own category for the primary variable is given by the sum of all coefficients holding only these indices except the first, which are ( $c_2, c_3, \dots, c_N$ ). To compare the performance of these two methods, we specify the null hypothesis that these two cell means are equal, or in other words that their difference is zero. As the set of all coefficients of our own category is a subset of the coefficients of the competitor category, this difference equals the sum of all coefficients whose index  $c_1$  equals 1.

In a more general situation some secondary variables are in the reference category, while others are not. Theorem 4.2.23 now tells us, that the mean of a cell containing the competitor category is the sum of all coefficients that only contain indices of the non-reference-categories. From Theorem 4.2.23 we further know that the mean of a cell containing the competitor category the is the same as explained in the previous paragraph. The null hypothesis implies that the difference of the two cell means is zero. This difference is the sum of coefficients whose index  $c_1$  equals 1 and of coefficients with non-reference category indices.

**Example 4.2.25.** Remember Example 4.2.24. Herein define  $X_1$  to be the primary category  $P_1$  and  $X_2, X_3$  to be  $S_2$  and  $S_3$ . Let us again look at ( $s_2^1, s_3^1$ ). As seen in Equation (4.2.6), the competitor effect is

$$\hat{\beta}_0 + \hat{\beta}_{1,1} + \hat{\beta}_{2,1} + \hat{\beta}_{3,1} + \hat{\beta}_{1,1;2,1} + \hat{\beta}_{1,1;3,1} + \hat{\beta}_{2,1;3,1} + \hat{\beta}_{1,1;2,1;3,1},$$

while our own effect is

$$\hat{\beta}_0 + \hat{\beta}_{2,1} + \hat{\beta}_{3,1} + \hat{\beta}_{2,1;3,1}.$$

Their difference is hence

$$\hat{\beta}_{1,1} + \hat{\beta}_{1,1;2,1} + \hat{\beta}_{1,1;3,1} + \hat{\beta}_{1,1;2,1;3,1}$$

which is exactly the sum of those coefficients that contain index  $c_1 = 1$  and the other non-reference-categories. Again, the intercepts cancel out.

### 4.2.3 Complete Null Hypothesis

We now wish to investigate total effects. To this end we compare the means of all cells containing the own category to the means of cells containing categories of all competitors. This can be achieved by simply adding up the means of individual cells. Alternatively, one could add all cell means containing the competitor category on one side, and add up all cell means containing the own category on the other side. Then the null hypothesis implies the equality of these two sums.

The main work has already been done. From Section 4.2.2, we know that all the terms appearing in the null hypotheses must carry the index  $c_1 = 1$ . All that remains is to determine the weight of each term in the linear combination representing the hypothesis.

Every coefficient with  $c_1 = 1$  symbolized as  $\hat{\beta}_{1,1;k_1,c_{k_1};\dots;k_r,c_{k_r}}$  describes its own cell it appears in while all the other variables are in the reference category. Coefficient  $\hat{\beta}_{1,1;k_1,c_{k_1};\dots;k_r,c_{k_r}}$  appears once in the cell

$$(p_1^1, s_2^0, \dots, s_{k_1-1}^0, s_{k_1}^{c_{k_1}}, s_{k_1+1}^0, \dots, s_{k_r-1}^0, s_{k_r}^{c_{k_r}}, s_{k_r+1}^0, \dots, s_N^0)$$

But such a coefficient also appears in each cell in which additional secondary variables are ‘raised’ from the reference category to a non-reference category:

$$(p_1^1, s_2^0, \dots, s_{k_1-1}^1, s_{k_1}^{c_{k_1}}, s_{k_1+1}^1, \dots, s_{k_r-1}^1, s_{k_r}^{c_{k_r}}, s_{k_r+1}^0, \dots, s_N^1).$$

Therefore such a coefficient will also appear as many times, as there are ‘higher’ options across all reference categories contained in its index. Let  $I$  be the set of indices which are in their reference category, i.e., which aren’t  $(k_1, \dots, k_r)$ . The total number of appearances of each  $c_1 = 1$  coefficient therefore is:

$$\begin{aligned} & 1 \text{ [its own category]} \\ & + \sum_{i \in I} (n_i - 1) \text{ [a change in 1 variable } X_i] \\ & + \sum_{i > j \in I} (n_i - 1)(n_j - 1) \text{ [changing 2]} \\ & + \sum_{i > j > k \in I} (n_i - 1)(n_j - 1)(n_k - 1) \text{ [changing 3]} \\ & + \dots \text{ changing 4 and up to } \#I - 1 \text{ variables} \\ & + \prod_{i \in I} (n_i - 1) \text{ [changing all variables belonging to } I] \\ & = \sum_{J \subset I} \prod_{j \in J} (n_j - 1) = \prod_{i \in I} n_i \end{aligned}$$

as can be seen in Lemma 4.5.5.

Consequently the total number of appearances of such a coefficient equals the product of the number of categories for the variables which according to the coefficient are in the reference category.

### 4.3 Application

The data for this example is taken from Gahler and Hruschka (2018a).

In Gahler and Hruschka (2018a), four procedures for allocating advertising budgets are compared by means of a simulation study. Two primary variables were considered independently, though we will only focus on one, the obtained returns after 40 periods. Furthermore, five secondary variables (functional form of the response function, disturbance level, saturation level, elasticity level and budget level) were considered, but, for brevity, we will consider only a subset of this data.

Most accurately, the following factor levels were eliminated: The third rule of thumb (ROT) in the primary variable, the multiplicative- and ADBUDG-response functions, the ‘similar’ saturation level and the low budget level.

This leaves us with three levels in the primary variable (the described algorithm and the first two rules of thumb), a secondary variable with two factor-levels (similar and varied elasticity levels) and a secondary variable with three factor-levels (three disturbance levels). Stated differently, we have

$$X_1 = P_1 = \text{Procedures} : (p_1^0 = \text{‘algorithm’}, p_1^1 = \text{‘first ROT’}, p_1^2 = \text{‘second ROT’}, )$$

$$X_2 = S_2 = \text{Elasticities} : (s_2^0 = \text{‘similar’}, s_2^1 = \text{‘varied’})$$

$$X_3 = S_3 = \text{Disturbances} : (s_3^0 = \text{‘low’}, s_3^1 = \text{‘moderate’}, s_3^2 = \text{‘high’})$$

The corresponding main-effect- (PUR1), pairwise-interaction- (PAIR1), and full-interaction-regression FULL1 models have respective  $R^2$ -values of 0.31, 0.58 and 0.65, indicating that the full interaction model is the best for explaining variance. Furthermore, the F-test conducted via anova (PUR1, PAIR1, FULL1)

confirms that the full model is to be preferred to its two nested models, with p-values beyond  $10^{-10}$ .

Note that this is a property that holds for the *entire* dataset with all six variables and sextuple interactions. We obtain our coefficient vector through summary(FULL1)

If we want, for example, to determine the effects of the algorithm and the first rule of thumb with varied elasticities in and high disturbances, these effects are represented by means of cells  $(p_1^0, s_2^1, s_3^2)$  and  $(p_1^1, s_2^1, s_3^2)$ , respectively. By Theorems 4.2.23 and 4.2.22, the former equals the sum of all coefficients holding indices (2,1) and (3,2), and the latter equals the sum of all coefficients holding indices (1,1), (2,1) and (3,2). Then the null hypothesis is:

$$\hat{\beta}_0 + \hat{\beta}_{2,1} + \hat{\beta}_{3,2} + \hat{\beta}_{2,1;3,2} = \hat{\beta}_0 + \hat{\beta}_{1,1} + \hat{\beta}_{2,1} + \hat{\beta}_{3,2} + \hat{\beta}_{1,1;2,1} + \hat{\beta}_{1,1;3,2} + \hat{\beta}_{2,1;3,2} + \hat{\beta}_{1,1;2,1;3,2}$$

which is equivalent to the difference being zero, and confirming the result of Section 4.2.2, only coefficients remain whose index  $c_1$  equals 1 (symbolized by 1, 1):

$$\hat{\beta}_{1,1} + \hat{\beta}_{1,1;2,1} + \hat{\beta}_{1,1;3,2} + \hat{\beta}_{1,1;2,1;3,2} = 0.$$

Calculating the right hand side gives a value of  $-0.0212 - 0.0199 + 0.0050 + 0.0164 = -0.0197$ , implying that the algorithm

Table 4.6: summary (FULL1)

Coefficient	Description	OLS Estimate
$\hat{\beta}_0$	Intercept	0.97678
$\hat{\beta}_{1,1}$	first ROT	-0.0212
$\hat{\beta}_{1,2}$	second ROT	0.0156
$\hat{\beta}_{2,1}$	varied	-0.0303
$\hat{\beta}_{3,1}$	moderate	-0.0119
$\hat{\beta}_{3,2}$	high	-0.0291
$\hat{\beta}_{1,1;2,1}$	first ROT:varied	-0.0199
$\hat{\beta}_{1,2;2,1}$	second ROT:varied	-0.0096
$\hat{\beta}_{1,1;3,1}$	first ROT:moderate	-0.0036
$\hat{\beta}_{1,2;3,1}$	second ROT:moderate	-0.0149
$\hat{\beta}_{1,1;3,2}$	first ROT:high	0.0050
$\hat{\beta}_{1,2;3,2}$	second ROT:high	-0.1031
$\hat{\beta}_{2,1;3,1}$	varied:moderate	0.0047
$\hat{\beta}_{2,1;3,2}$	varied:high	0.0100
$\hat{\beta}_{1,1;2,1;3,1}$	first ROT:varied:moderate	0.0175
$\hat{\beta}_{1,2;2,1;3,1}$	second ROT:varied:moderate	-0.0595
$\hat{\beta}_{1,1;2,1;3,2}$	first ROT:varied:high	0.0164
$\hat{\beta}_{1,2;2,1;3,2}$	second ROT:varied:high	-0.3024

has higher returns than the ROT. The F-test on the linear restriction defined by vector  $v = (0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0)$  gives a p-value  $> 0.4$  Greene (1993), signalling that they are not significantly higher.

Interpretation is straightforward: Under the condition tested (varied elasticities and high disturbances) we cannot conclude that the algorithm yields higher sales than the first ROT.

Now for the main question: Does one procedure outperform the other *overall*?

The triple interaction term  $\hat{\beta}_{1,1;2,1;3,2}$  appears only in its own right, representing the secondary categories  $(s_2^1, s_3^2)$  and hence has a weight of 1.

The term  $\hat{\beta}_{1,1;3,2}$  appears in its own cell as it represents category  $(s_3^2)$ , but also appears in cell  $(s_2^1, s_3^2)$  and hence has a weight of 2.

Likewise,  $\hat{\beta}_{1,1;2,1}$  appears in its own cell as well as in cells  $(s_2^1, s_3^2)$  and  $(s_2^1, s_3^1)$ , and hence has a weight of 3.

Finally,  $\hat{\beta}_{1,1}$  appears in its own cell and 5 more cases, leading to a weight of 6. Note that these results are in agreement with Section 4.2.3. There we show that we only have to consider coefficients having 1, 1 in the subscript. We also demonstrate in this section that the weight of such a coefficient corresponds to the product of the number of categories for the reference categories indicated by this coefficient.

The null hypothesis comparing the two procedures is represented by

$$v' = (0, 6, 0, 0, 0, 0, 3, 0, 0, 0, 2, 0, 0, 0, 0, 0, 1, 0)$$

which holds the weight explained and leads to

$$6\hat{\beta}_{1,1} + 3\hat{\beta}_{1,1;2,1} + 2\hat{\beta}_{1,1;3,2} + \hat{\beta}_{1,1;2,1;3,2} = 0$$

The left hand side of this expression sums up to -0.161. The F-test on the linear restriction defined by vector  $v'$  results in a p-value  $p < 0.05$  which confirms that overall the algorithm outperforms the first rule of thumb.

## 4.4 Conclusion

The developed method allows researchers to use full ANOVA models with many independent variables in different application areas if interest focuses on the effects of categories of one selected independent variable. Full ANOVA models are appropriate if highest-order interactions are significant and of considerable size. The method might be extended in at least three respects.

Firstly, it could be adapted to ANOVA models with many variables which contain some, but not all higher interactions terms.

The second possible extension is connected with the fact that the method is restricted to categorical variables. How to deal with continuous variables and their interactions is another interesting, albeit challenging research question.

Thirdly, we know from Lemma 4.5.6 that the results presented only have a well defined OLS-estimator if the design matrix contains every possible row. If it does not, several actions for adjustment may be taken, from increasing the sample size, over merging categories to eliminating entire variables. But perhaps there is a mathematical approach yet to be found for the case of ill-defined OLS-estimators that still gives us the correct linear combinations of coefficients.

## 4.5 Appendix A: Proofs

**Lemma 4.5.1.** *Some linear algebra frequently used:*

- 1.) For  $m, n, k \in \mathbb{N}^+$  let  $A$  be a  $m$  by  $n$  matrix and  $B$  be an  $n$  by  $k$  matrix, then  $(AB)^T = B^T A^T$ .
- 2.) Let  $A$  be an invertible matrix, then  $(A^T)^{-1} = (A^{-1})^T$ .
- 3.) Let  $A, B$  be invertible matrices of the same dimension. Then  $(AB)^{-1} = B^{-1} A^{-1}$ .
- 4.) Let  $P$  be a permutation matrix. Then  $P$  is orthogonal, i.e.  $P^T = P^{-1}$ .

*Proof.* All taken from Abadir and Magnus (2005)

- 1.) Abadir and Magnus (2005) P.26, Ex.2.15 (a)
- 2.) Abadir and Magnus (2005) P.26, Ex.2.15 (c)
- 3.) Abadir and Magnus (2005) P.26, Ex.2.15 (d)
- 4.) Abadir and Magnus (2005) P.32, Ex.2.29 (d)

□

**Lemma 4.5.2.** *Let  $R, V, X$  be as in Proposition 4.2.2. Then  $V$  is its own inverse and  $V$  (and its transpose) and  $R$  uniquely solve the equations*

$$VX^T = X^T R \quad (4.5.1)$$

$$R^T X = XV^T \quad (4.5.2)$$

*Proof.* The second equation follows from the first, since

$$R^T X = ((R^T X)^T)^T = (X^T R)^T = (VX^T)^T = XV^T. \quad (4.5.3)$$

As  $X^T$  is invertible, the unique solution is  $V = X^T R (X^T)^{-1}$ , which is exactly the matrix we defined as  $V$  in Section 4.2.1, and finally

$$V^2 = X^T R (X^T)^{-1} X^T R (X^T)^{-1} = X^T R R (X^T)^{-1} = X^T (X^T)^{-1} = \mathbb{1}. \quad (4.5.4)$$

□

The same calculations go through in the later sections with  $V$  and  $R_{\text{red}}$ , since  $R_{\text{red}}$  is symmetric and orthogonal.

**Lemma 4.5.3.** *Let  $X$  be a reduced design matrix. Then  $X$  has full rank, and is hence, invertible.*



*Proof.* The way we constructed the matrix  $X$  columnwise, the first 1 that appears is in the first row where it is measured (though there might be more 1s below). Up to this point, counting from top to bottom, this is the last non-zero entry in the row it appears in, as a 1 to the right of it would represent a higher interaction effect in which it appears. We can therefore define a permutation matrix  $P$  which, multiplied by  $X$  gives  $X' = PX$  as a lower triangular matrix with 1s on the main diagonal and only zeroes to the right of this diagonal. We know that the determinant of a lower triangular matrix is the product of its diagonal entries, hence  $\det(X') = 1$ . We know furthermore, that the determinant of a permutation matrix is its sign, hence  $\det(P) \in \{-1, 1\}$ . We can therefore conclude

$$\det(X) = \det(P^{-1}X') = \det(P^{-1})\det(X') = \frac{1}{\det(P)} \in \{-1, 1\}. \quad (4.5.5)$$

In particular,  $\det(X) \neq 0$  and therefore  $X$  is invertible and has full rank.  $\square$

**Lemma 4.5.4.** *Let  $R_0, R$  be as in the two construction steps throughout Section 4.2.1. For indexing reasons, we write  $\Phi$  instead of  $R_0$ .*

*As  $\Phi$  is symmetric,  $v_i := \sum_{k=1}^n \Phi_{ik} = \sum_{k=1}^n \Phi_{ki}$  and the thereby defined vector  $v$  contains the row-wise (and column-wise) sums of entries of  $\Phi$ .*

1.)  $\Phi$  has the following property:

$$\text{for all } i, j \in 1, \dots, n : \Phi_{ij} \sum_l \frac{\Phi_{jl}}{v_l} = \Phi_{ij} \frac{v_j}{v_i}. \quad (4.5.6)$$

2.) There is a diagonal matrix  $D$  such that  $R^T D = R$ .

3.) It further holds that  $D^{-1} R^T = R$ .

4.) The product  $R^T y$  replaces the entries of a dependent variable  $y$  belonging to the first usual row with the mean of the entries belonging to the fourth usual row and vice versa. It further replaces the entries belonging to the second and third usual row with their respective mean.

*Proof.* 1.)  $\Phi$  is symmetric with entries in  $\{0, 1\}$ , so the statement is trivial for  $\Phi_{ij} = 0$ .

Now let  $\Phi_{ij} = 1$ . By construction (see step 1) this can mean one of three things:

a)  $i \in I_{i_1}, j \in I_{i_2}$

b)  $j \in I_{i_1}, i \in I_{i_2}$

c)  $\exists k \in I \setminus \{i_1, i_2\}, i, j \in I_k$

We begin with the simple situation c).

Let  $k$  be such that  $i_1 \neq k \neq i_2$  and  $i, j \in I_k$ . In particular, this means that  $v_i = v_j = \#I_k$ . Furthermore, for every  $l$  such that  $\Phi_{jl} \neq 0, v_l = v_j$ , therefore

$$\sum_l \frac{\Phi_{jl}}{v_l} = \frac{1}{v_j} \sum_l \Phi_{jl} = \frac{1}{v_j} v_j = 1 = \frac{v_j}{v_i}. \quad (4.5.7)$$

We now prove the statement for a), and b) follows immediately, since the statement is symmetric in  $i, j$  and  $i_1, i_2$ .

Given  $i_1, i_2$ , let  $i \in I_{i_1}, j \in I_{i_2}$ , such that  $\Phi_{ij} \neq 0$ . As column  $j$  has a 1 in every entry from  $I_{i_1}$ , we know that  $v_j = \#I_{i_1}$ . This happens in *every* row  $l$  with  $l \in I_{i_1}$ , therefore in these rows (and hence also row  $i$ ) the sum of entries is  $v_l = \#I_{i_2}$  and hence

$$\sum_l \frac{\Phi_{jl}}{v_l} = \frac{1}{\#I_{i_2}} \sum_l \Phi_{jl} = \frac{\#I_{i_1}}{\#I_{i_2}} = \frac{v_j}{v_i}. \quad (4.5.8)$$

Note, that the equation  $\sum_l \frac{\Phi_{jl}}{v_l} = \frac{v_j}{v_i}$  does *not* hold in general! But it does hold for  $i, j$  such that  $\Phi_{ij} = 1$ , which is exactly what we have shown. Therefore the equation  $\Phi_{ij} \sum_l \frac{\Phi_{jl}}{v_l} = \Phi_{ij} \frac{v_j}{v_i}$  *does* hold for all  $i$  and  $j$ .

2.) Define  $\Delta$  to be the diagonal matrix containing the elements of  $v$  in order, i.e.:

$$\Delta_{ij} = \begin{cases} v_i & \text{if } i = j \\ 0 & \text{else} \end{cases} \quad (4.5.9)$$

By construction, we know that  $R = \Phi \Delta^{-1}$  and  $R^T = (\Phi \Delta^{-1})^T = \Delta^{-1} \Phi$ .

We now define  $\eta_i := 1/(\sum_{k=1}^n R_{ik}) = 1/(\sum_{k=1}^n R_{ki}^T)$  to be the inverse of the row-wise sum of  $R$ , which is the same as the inverse of the column-wise sum of  $R^T$ .

We further set  $D$  to be the diagonal matrix containing the elements of  $\eta$  in order, i.e.:

$$D_{ij} = \begin{cases} \eta_i & \text{if } i = j \\ 0 & \text{else} \end{cases} \quad (4.5.10)$$

We need to prove  $R^T D = R$  by showing that their entries are equal:

$$(R^T D)_{ij} = \sum_k R_{ik}^T D_{kj} \quad (4.5.11)$$

$$= \sum_k R_{ki} D_{kj} \quad (4.5.12)$$

$$= R_{ji} D_{jj} \quad (4.5.13)$$

$$= (\Phi \Delta^{-1})_{ji} \eta_j \quad (4.5.14)$$

$$= \sum_k \Phi_{jk} \Delta_{ki}^{-1} \frac{1}{\sum_l R_{jl}} \quad (4.5.15)$$

$$= \Phi_{ji} \Delta_{ii}^{-1} \frac{1}{\sum_l (\Phi \Delta^{-1})_{jl}} \quad (4.5.16)$$

$$= \frac{\Phi_{ji}}{v_i} \frac{1}{\sum_l (\sum_k \Phi_{jk} \Delta_{kl}^{-1})} \quad (4.5.17)$$

$$= \frac{\Phi_{ji}}{v_i} \frac{1}{\sum_l \Phi_{jl} \Delta_{ll}^{-1}} \quad (4.5.18)$$

$$= \frac{\Phi_{ji}}{v_i} \frac{1}{\sum_l \frac{\Phi_{jl}}{v_l}} \quad (4.5.19)$$

Remember from Equation (4.5.6) that  $\Phi_{ij} \sum_l \frac{\Phi_{jl}}{v_l} = \Phi_{ij} \frac{v_j}{v_i}$ . We can therefore conclude:

$$\frac{\Phi_{ji}}{v_i} \frac{1}{\sum_l \frac{\Phi_{jl}}{v_l}} \stackrel{(4.5.6)}{=} \frac{\Phi_{ji}}{v_i} \frac{1}{\frac{v_j}{v_i}} \quad (4.5.20)$$

$$= \frac{\Phi_{ji}}{v_i} \frac{v_i}{v_j} \quad (4.5.21)$$

$$= \frac{\Phi_{ji}}{v_j} \quad (4.5.22)$$

and conversely

$$R_{ij} = (\Phi \Delta^{-1})_{ij} \quad (4.5.23)$$

$$= \sum_k \Phi_{ik} \Delta_{kj}^{-1} \quad (4.5.24)$$

$$= \Phi_{ij} \Delta_{jj}^{-1} \quad (4.5.25)$$

$$= 1/v_j \Phi_{ij} = \frac{\Phi_{ij}}{v_j} \quad (4.5.26)$$

$$= \frac{\Phi_{ji}}{v_i} \stackrel{\text{above}}{=} (R^T D)_{ij} \quad (4.5.27)$$

and therefore  $R = R^T D$ .

3.)

$$D^{-1} R^T = D^{-1} (R^T D)^T = D^{-1} D^T (R^T)^T = D^{-1} D R = R \quad (4.5.28)$$

4.) Remember the definitions of  $I_1, I_4$  and keep in mind  $R^T = \Delta^{-1} \Phi$ .

Now, for every index  $l \in I_1$  we have

$$(R^T y)_l = (\Delta^{-1} \Phi y)_l = \sum_{k \in I} (\Delta^{-1} \Phi)_{lk} y_k = \sum_{k \in I} \left( \sum_h \Delta_{lh}^{-1} \Phi_{hk} \right) y_k = \quad (4.5.29)$$

$$= \sum_{k \in I} (\Delta_{ll}^{-1} \Phi_{lk}) y_k = \sum_{k \in I} \frac{\Phi_{lk}}{v_l} y_k. \quad (4.5.30)$$

Since  $l \in I_1$ , by construction,  $\Phi_{lk} = 1 \iff k \in I_4$ , and otherwise  $\Phi_{lk} = 0$ , and therefore  $v_l = \#I_4$  and

$$\sum_{k \in I} \frac{\Phi_{lk}}{v_l} y_k = \sum_{k \in I_4} \frac{\Phi_{lk}}{v_l} y_k = \sum_{k \in I_4} \frac{y_k}{\#I_4}, \quad (4.5.31)$$

which is precisely the mean of entries of  $y$  belonging to  $I_4$ . The other parts of the statement follow analogously.  $\square$

**Lemma 4.5.5.**

$$\sum_{I \subseteq \{1, \dots, N\}} \prod_{i \in I} (n_i - 1) = \prod_{i=1}^N n_i \quad (4.5.32)$$

*Proof.* By complete induction on  $N$ .

Basis:  $N = 2 : \{1, 2\}$  has 4 subsets:  $\emptyset, \{1\}, \{2\}$  and  $\{1, 2\}$ . Remember that the product over an empty index set is 1. Then we have

$$\sum_{I \subseteq \{1, 2\}} \prod_{i \in I} (n_i - 1) = \prod_{i \in \emptyset} (n_i - 1) + \prod_{i \in \{1\}} (n_i - 1) + \prod_{i \in \{2\}} (n_i - 1) + \prod_{i \in \{1, 2\}} (n_i - 1) \quad (4.5.33)$$

$$= 1 + (n_1 - 1) + (n_2 - 1) + (n_1 - 1)(n_2 - 1) = n_1 n_2, \quad (4.5.34)$$

where the last equality was already seen in Equation (4.2.4)

Step:  $N - 1 \rightarrow N$

Let the statement be known for  $N - 1$ . We split the index set into those that contain 1 and those that do not, and then use the induction step on both sets.

$$\begin{aligned} \sum_{I \subseteq \{1, \dots, N\}} \prod_{i \in I} (n_i - 1) &= \sum_{1 \in I \subseteq \{1, \dots, N\}} \prod_{i \in I} (n_i - 1) + \sum_{1 \notin I \subseteq \{1, \dots, N\}} \prod_{i \in I} (n_i - 1) \quad (4.5.35) \\ &= (n_1 - 1) \sum_{1 \in I \subseteq \{1, \dots, N\}} \prod_{i \in I, i \neq 1} (n_i - 1) + \sum_{I \subseteq \{2, \dots, N\}} \prod_{i \in I} (n_i - 1) \quad (4.5.36) \end{aligned}$$

$$= (n_1 - 1) \sum_{I \cup \{1\}, I \subseteq \{2, \dots, N\}} \prod_{i \in I} (n_i - 1) + \prod_{i=2}^N n_i \quad (4.5.37)$$

$$= (n_1 - 1) \sum_{I \subseteq \{2, \dots, N\}} \prod_{i \in I} (n_i - 1) + \prod_{i=2}^N n_i \quad (4.5.38)$$

$$= (n_1 - 1) \prod_{i=2}^N n_i + \prod_{i=2}^N n_i = \prod_{i=1}^N n_i \quad (4.5.39)$$

□

**Lemma 4.5.6.** *Let  $X$  be a reduced design matrix of dimension  $\tilde{N}$  and let  $M$  be an  $m \times \tilde{N}$  matrix consisting of the rows of  $X$ . Then  $M^T M$  is invertible if and only if  $M$  is a generic design matrix (i.e.  $M$  contains every row of  $X$  at least once).*

*Proof.* ‘ $\Leftarrow$ ’ Let  $M$  be a generic design matrix. Since every row of  $X$  appears as a row in  $M$ , we can find an  $m \times m$  permutation matrix  $P$  such that the first  $\tilde{N}$  rows of  $PM$  are the exact rows of  $X$  in order. Since  $X$  is invertible by Lemma 4.5.3,  $X$  has rank  $\tilde{N}$  and so does  $PM$ . Since permuting rows does not change the rank (see Greene (1993) P. 18-19 2-43 & 2-45),  $M$  has full rank  $\tilde{N}$  as well. Therefore,  $M^T M$  is a square matrix of dimension  $\tilde{N}$  with rank  $\tilde{N}$  (see Greene (1993) P. 19 2-46) and hence invertible.

‘ $\Rightarrow$ ’ By contraposition. Assume that  $M$  is not a generic design matrix. Then there is a row of  $X$  which does not appear in  $M$ .

We will first show that  $M$  does not have rank  $\tilde{N}$ , so without loss of generality, we may assume that  $M$  has at least  $\tilde{N}$  rows. If it has fewer, that particular statement must already be true.

So let the length of  $M$  be at least  $\tilde{N}$ . Define a permutation matrix  $P$  such that all the unique rows of  $M$  are in the first  $\tilde{N}$  rows. As there is a row of  $X$  that now does not appear in the first  $\tilde{N}$  rows of  $PM$ , at least one of them must appear twice, since  $M$  only consists of rows of  $X$ . As all rows below row  $\tilde{N}$  are non-unique,  $PM$  has at most  $\tilde{N}-1$  unique rows, and hence  $\text{rank}(PM) < \tilde{N}$ . Again, the same holds for  $M$ , as permutations do not change the rank.

Finally,  $M^T M$  is a square matrix of dimension  $\tilde{N}$  with rank  $< \tilde{N}$  and hence cannot be invertible.

In particular, all OLS-estimators appearing in this article are well-defined.  $\square$

## 4.6 Appendix B: Further Comments

In this section we give further explanations to Definition 4.2.21 and the proofs of Theorem 4.2.16, Lemma 4.5.3, 4.5.4 and 4.5.5. It was originally intended as an audioscript for a journal and is supposed to facilitate the the individual steps of the definition and the proofs. It is therefore written in very plain terms, as the formal versions are already written above, in the corresponding sections.

**Definition 4.2.21** We now define the generic design matrix in its most general form.

Note that this definition is a generalization of Definition 4.2.4 and is just the general definition of a designmatrix for dummy-coded variables, in which each case appears exactly once.

It can be helpful to compare this construction to the matrix from Example 4.2.24

In Lemma 4.5.5 in 4.5 Appendix A, we will prove that it is a square matrix, and in Lemma 4.5.3 we will prove that it is always invertible.

We start with  $N$  dummy-coded variables, which we call  $X_1, \dots, X_N$ , and each of these can take only one of a finite number of states. We call these states the categories, so for  $X_i$ , we say it has  $n_i$  categories. The total number of categories is  $\tilde{N}$ , which is the product over all cardinalities of categories. Therefore, this will be the number of rows and columns of the reduced design matrix.

We will index the rows by the cells, and the columns by the coefficients.

So for every row, the row index will be the vector of the cases each variable is in, written as  $x_i^{k_i}$ . This superscript is just a notation, not an exponent, and signifies, just like before, that variable  $x_i$  takes the category  $k_i$ .

For understanding the indexing of the rows, it is useful to keep the example in mind, where the number of categories of each variable is 10. That way, the superscript in sequence represent the digits of a decimal integer in reverse order. That number will be  $k$  in row number  $k + 1$ . In row one, this number will be zero, so the row index has a zero in every superscript. In row 2, the first variable will take the first non-reference category, and all others will be zero, so the superscripts will be 1 followed by zeroes, which is  $k = 0 \dots 0001 = 1$  in reverse, since  $k + 1 = 2$ .

In the decimal system, to get the least significant digit of a number, we take the remainder modulo 10. For the second least significant digit, we subtract the least significant from the number, divide by 10, and again take the remainder modulo 10.

We use this very construction to index the rows. In row  $k + 1$ , the category of  $x_1$  will just be  $k$  modulo the size of the variable, hence modulo  $n_1$ . For the next variable,

we subtract the previous term and divide by  $n_1$ , then take the remainder mod  $n_2$ . We continue this process for all variables. Keeping this construction in mind will be very helpful later on for identifying the correct rows for all cases.

We now construct the columns. Every column will be indexed as the Greek letter beta with the configuration of all variables that are not in the reference category in the subscript. In the first column, representing the intercept, all variables are in their reference category, so we denote that simply by  $\beta_0$ .

The columns are indexed from left to right by number of variables per interaction term, starting with 0, where we have the intercept. Next, with one variable, so just the main effects, we start counting up the non-reference categories of the first variable  $X_1$  which will be  $\beta_{1,1}$  up to  $\beta_{1,n_1-1}$ . For the main effects, the beta-subscript is  $(i, j)$  for the main effect of  $X_i$  in non-reference category  $x_i^j$ . The general subscript notation is as follows:  $k_1, t_1; k_2, t_2; \dots; k_l, t_l$  is the subscript of an interaction of  $l$  variables  $X_{k_1}, \dots, X_{k_l}$  which are in the non-reference-categories  $x_{k_1}^{t_1}, \dots, x_{k_l}^{t_l}$ .

They are arranged from left to right in  $N + 1$  clusters of coefficients with the same number  $l$  of interacting variables. Within each cluster they are sorted by variables, so cluster  $l$  is divided into subclusters indexed  $(k_1, \dots, k_l)$  which are sorted within the cluster  $l$  by incrementing from left to right. Within the subcluster  $(k_1, \dots, k_l)$ , the coefficients are sorted by incrementing the categories  $(t_1, \dots, t_l)$  from left to right.

We have now indexed the rows of the matrix by cases, and the columns by coefficients.

All that's left to be done is fill the matrix with values. We do so by simply assigning a 1 to each pair  $(k + 1, j)$  of row  $k + 1 = (k_1, \dots, k_N)$  and column  $j = \beta_{(i_1, a_1; \dots, i_l, a_l)}$  when all the categories that appear in the  $\beta$ -subscript are met by the cases in the columns. So for each pair  $(i_p, a_p)$  in the subscript of  $\beta$ , we want that the variable  $X_{i_p}$  takes the category  $a_p$ , which means that the row index contains the term  $x_{i_p}^{a_p}$ . If this is true for all categories appearing in the subscript of beta, the entry of the matrix  $X$  in row  $k + 1$  and column  $j$  will contain a 1, and if not, a 0.

We then calculate the number of coefficients, and we see with the help of Lemma 4.5.5 from the 4.5 Appendix A that that is equal to the number of rows, and hence square, and it will be invertible which is proven in Lemma 4.5.3.

**Theorem 4.2.16** In this theorem we condense all the work we have done previously in order to get the result we need. We are in the case of 2 variables with 2 cases each, and a generic design matrix  $M$ .

The claim is, that the permutation matrix  $R^T$ , which changes the reference case in both variables, can be pulled through the OLS estimator, and comes out on the other end

as  $V^T$ . This means that the effect of the two non-reference categories can be determined by multiplying  $V^T$  with the original OLS estimator.

It is a generalization of Proposition 4.2.8, which itself is a generalization of Proposition 4.2.2. Therefore, if there are problems with certain steps, we recommend analyzing the proofs of those propositions. Furthermore we use Proposition 4.2.12, which is mainly proven with the help of Lemma 4.5.4 in 4.5 Appendix A.

We also use some linear algebra, found in Lemma 4.5.1 of 4.5 Appendix A. This is, however, elementary, as it is simply the transpose and inverse of a product of matrices, and the inverse of the transpose.

Now, for the proof.

We start with the left hand side, and in the first step we write out the OLS of  $M$  with  $y$ . As  $V^T$  is invertible, we can pull its inverse into the inverse product, on the right hand side. This is easier seen by going backwards, as the inverse of a product is the product of the inverses in switched order.

Going to the next line, we know from Lemma 4.5.2 that  $V^2$  is the identity matrix, so  $V$  is its own inverse, and hence,  $V^T$  is its own inverse as well. The next step is a consequence of Proposition 4.2.12.3), we get from transposing this equation, that  $R^T M$  is equal to  $MV^T$ .

To the next line, we know from Proposition 4.2.12.4) that there is a diagonal matrix such that  $R = R^T D$ , or equivalently,  $R^T = RD^{-1}$ . Again, we call pull  $R$  through  $M^T$ , and receive  $V$  on the other side.

To the next line, we know from Proposition 4.2.12.5) that there is a matrix  $V_2$  such that  $M^T D^{-1} = V_2 M^T$ . Both  $V$  and  $V_2$  are invertible, so we can pull them out of the left side of the bracket, by multiplying their inverses from the right side of the bracket.

To the next line,  $V$  is its own inverse, and can be pulled through  $M^T$  as  $R$ .

To the next line,  $R$  can be written as  $D^{-1} R^T$ , and  $D^{-1}$  can be pulled BACK through  $M^T$  as  $V_2$ , as seen before. Finally,  $V_2$  and its inverse cancel out, and we get the desired term.

**Lemma 4.5.3** To understand this proof it is absolutely necessary to understand the construction of  $X$ , i.e. Definition 4.2.21.

It is again helpful to look at the matrix in Example 4.2.24.

A reformulation of the proof is the following. Take any column. Want to show that the first "1" that it appears in will be in the row in which it is the last one. So if you



follow down any column to its first appearing "1", there will only be zeroes to the right of that particular "1".

Why is that true, and how does that help us?

It is obviously true for the intercept column: the very first entry is a "1", representing the reference case in every variable. Therefore in this row, no non-reference cases appear. Naturally that means that there can't be another column containing a "1" in this row, because for every other column there is a one in each row of the designated non-reference-cases. But those simply do not exist in the first row, where all variables are in the reference category.

Now, take any column indexed by  $\beta_{i_1, a_1; \dots, i_l, a_l}$ . In this column, there is a "1" in exactly those rows  $(x_1^{k_1}, \dots, x_N^{k_N})$ , where  $k_{i_m} = a_{i_m}$  for all  $m = 1, \dots, l$ , so in every non-reference case described by the beta subscript. The very first appearance is the one where all non-specified cases are zero, i.e.

$$(x_1^0, \dots, x_{i_1-1}^0, x_{i_1}^{a_1}, x_{i_1+1}^0, \dots, x_{i_l-1}^0, x_{i_l}^{a_l}, x_{i_l+1}^0, \dots, x_N^0)$$

in the shorter version, the row index is

$$(0, \dots, 0, a_1, 0, \dots, 0, a_l, 0, \dots, 0)$$

.

We show by contradiction that there is no further 1 to the right of it:

Assume that there is another 1 to the right of it. Because of the ordering imposed on the columns in the definition this can mean that either the pre-specified variables are in a different category (which is impossible, because they are in the row with the pre-specified values) or that another variable is raised from a reference category into a non-reference category (which is impossible too, since by Definition 4.2.21 a "1" means that all non-reference-categories that appear in the beta-subscript must be in these cases in the corresponding row, but these are all zero), or both, which is impossible for both of the previous reasons.

Define a permutation as follows: for every column  $i = 1, \dots, N$  let  $p(i)$  be the row in which the first "1" appears. Because of the previous reasoning, this is a well-defined, bijective map from  $\{1, \dots, N\}$  to itself, hence a permutation. We then send row  $p(i)$  to row  $i$  [i.e. we define the inverse permutation] and multiply the corresponding permutation matrix  $P$  with  $X$ , so for every column  $i$ , the first appearing "1" will be in row  $i$  instead of row  $p(i)$ . In other words, it will be on the diagonal. As we only changed the rows on their own, they will keep the property that this (now diagonal) "1" is the last non-zero entry. So to the right of the diagonal there can only be zeroes. Therefore it is a lower triangular matrix with "1" in every diagonal entry.

The rest of the proof is trivial.

**Lemma 4.5.4** To understand this Lemma it is imperative to precisely follow the construction of  $R_0$  and  $R$ , so here is a quick recap.  $M$  is a generic design matrix of length  $n$  and  $a = (a_1, \dots, a_n)$  is its manifest vector. All numbers that appear in  $a$  are the numbers 1, 2, 3 and 4, each in its position corresponding to the usual rows of  $X$  in  $M$ . We define the four sets  $I_1, I_2, I_3$  and  $I_4$ , where  $I_j$  contains precisely those indices from 1 to  $n$  whose positions in the manifest vector are  $j$ .

$R_0$  and  $R$  are constructed such that the product  $Ra$  leaves the 2- and 3-positions constant, and changes the 1- and 4-positions.

For every column  $r^l$ , if  $l$  is in  $I_2$  or  $I_3$ , there will be ones in precisely those rows whose indices appear in  $I_2$  or  $I_3$  respectively. If  $l$  is in  $I_1$ , it will have ones in those rows whose indices appear in  $I_4$  and vice versa. We now call this matrix  $\Phi$ , because the subscript 0 of  $R_0$  would make things more complicated than necessary.

We prove the first part of the proposition. Lemma 4.5.4 is proven for the general situation, but we can just as well call  $i_1 = 1$  and  $i_2 = 4$  and these are the indices which we want to swap.

Now,  $v_i$  is defined as the sum of entries of row  $i$  of  $\Phi$ , and since  $\Phi$  is symmetric, it is equal to the sum of entries in column  $i$ . By construction of  $\Phi$ , every entry of  $v$  is non-zero.

We start with 1.c). Without loss of generality we say  $k = 2$ , so  $i, j$  are in  $I_2$ . Therefore the sum of elements of row  $i$  and likewise  $j$  is the cardinality of  $I_2$ , because precisely that many "1's" appear in these rows and columns. And that further holds for every  $l$  in  $I_2$ . The denominator is hence constant and independent of the summation index, and can be factored out. All other appearing terms are equal to the cardinality of  $I_2$ , and cancel out.

It remains to show a), so assume  $i$  is in  $I_1$  and  $j$  is in  $I_4$ , and  $\Phi_{ij}$  is 1. By construction, column  $j$  has a 1 in every row belonging to  $I_1$  so  $v_j = \#I_1$ . Likewise, every row  $l$  belonging to  $I_4$  has a 1 in every column from  $I_1$  so  $v_l = \#I_4$ . The rest is just as in the formula.

Part 2 is a long construction, but a straightforward proof, therefore comments will be on every new "="-sign.

- 1.) Definition of matrix multiplication
- 2.) Definition of the transpose
- 3.)  $D_{kj} = 0$  unless  $k = j$ , so only the  $k = j$ -case remains
- 4.) Construction of  $R$  by dividing  $\Phi$  by  $\Delta$ , also definition of  $D_{jj}$
- 5.) Definition of matrix multiplication, also definition of  $\eta_j$

- 6.)  $\Delta_{ki} = 0$  unless  $k = i$ , so only the  $k = i$ -case remains, also construction of  $R$  by dividing  $\Phi$  by  $\Delta$
  - 7.) Definition of  $\Delta_{ii}$ , also definition of matrix multiplication
  - 8.)  $\Delta_{kl} = 0$  unless  $k = l$ , so only the  $k = l$ -case remains
  - 9.) Definition of  $\Delta_{ll}$
  - 10.) Part 1 of the proposition
  - 11.) Inverse fraction
  - 12.) Cancellation
- and conversely

- 1.) Construction of  $R$  by dividing  $\Phi$  by  $\Delta$
- 2.) Definition of matrix multiplication
- 3.)  $\Delta_{kj} = 0$  unless  $k = j$  so only the  $k = j$ -case remains
- 4.) Definition of  $\Delta_{jj}$
- 5.) Pulling  $\Phi$  into the numerator
- 6.)  $\Phi$  is symmetric
- 7.) Steps 1. to 12. above, backwards.

Part 3 is straightforward, so it remains to show part 4.

We need to calculate  $(R^T y)_l$  for each of the 4 possibilities of  $l$ , but only  $l$  in  $I_1$  is shown, as all other cases follow by the same manner. For  $l$  in  $I_1$  we need to show that  $(R^T y)_l$  is the mean of all entries of  $y$  belonging to  $I_4$ . We explain along the "="-sings

- 1.) Equation in the first line of the proof
- 2.) Definition of matrix-vector multiplication
- 3.) Definition of matrix multiplication
- 4.)  $\Delta_{lh} = 0$  unless  $l = h$  so only that case remains
- 5.) Definition of  $\Delta_{ll}$
- 6.) As explained above, only those indices of  $I$  remain, which belong to  $I_4$ , since for all others the entry is 0.
- 7.) As explained above,  $v_l$  has  $\#I_4$  elements.

Comparing the very first and last terms gives us exactly the statement of part 4.

**Lemma 4.5.5** In this Lemma we prove that the reduced design matrix is always a square matrix. If parts of the statement or its proof are unclear, we recommend examining Definition 4.2.21, which is the definition of the reduced design matrix, and Example 4.2.24 where there is an example with 3 variables.

The reduced design matrix has in its columns the coding of the coefficients of the intercept, the main effects and the interaction coefficients. In its rows, the individual cells. The number of the individual cells is just the product over all of the category sizes. So we have  $N$  variables, and for each of these variables we have  $n_i$  cases, so the product over all these cases is just the number of rows of the reduced design matrix.

On the other hand we have the number of the coefficients which will be estimated. That can be calculated as follows: Since every variable will have one case as the reference case, there will be  $n_i - 1$  coefficients that can be calculated. And we take the intercept, we take all the main effects, which is just the sum of all these  $n_i - 1$ 's and then the double interactions, which are the double products of these numbers, triple interactions and so on. In total, for every subset of  $\{1, \dots, N\}$ , we take the product over all of these  $(n_i - 1)$  terms.

We prove the equality by complete induction over  $N$ .

We start with the basic case of  $N = 2$ , which means the set we consider is the set that consists of 1 and 2, which has 4 subsets: the empty sets, the singletons containing 1 and likewise 2, and finally the set that contains both. We then prove the formula by taking all the subsets and taking the products. Remember that the product over the empty set is 1, because there is nothing to be multiplied. That "1" will correspond to the intercept. The product over index sets with just one element is just the term itself, so we have  $1 + n_1 - 1 + n_2 - 1$  and finally the product. Now we can just multiply these out and rearrange them to get  $n_1 n_2$ , which can also be seen in Equation 4.

For the induction step we take the step from  $N - 1$  to  $N$ , so we assume that the statement is true for  $N - 1$ , and we want to show it for  $N$ . Now, we start on the left hand side of the equation and we split the sets over which we add up into 2 subsets, and these are the sets that include 1 in the index and those that don't and that's the first step.

In the next step, since the terms on the left hand side of the addition all include 1, we can factor out  $n_1 - 1$ , because  $i = 1$  is in all of these sets, and they are a factor in all of the summands, and by the law of distribution, we can factor them out of the sum.

Now, on the right hand side in the next step we see that we have the exact formula necessary for the induction step, and it has an index set of  $\{2, \dots, N\}$ , so that's  $N - 1$  elements, so we can perform the induction step there and we know that that is the product over all elements from 2 to  $n$ .

On the left hand side we just rewrote the index set, because if we sum over the set of the indices from 1 to  $N$  that contain 1, and then take the product over all that do not

contain 1, that is just the same as summing over those that do not contain 1. In the next step we see that this term does not depend on the index 1 anymore, so that it's just the sum over the subsets from 2 to  $N$ , and so just like before what we can apply the induction step and see that it's the product of  $i$  from 2 to  $N$  over  $n_i$ .

And now we have this very term multiplied by  $n_1 - 1$  and on itself, and if we factorize that, we see that's just  $n_1$  times the product of  $i$  from 2 to  $N$  over  $n_i$  so that's the product of  $i$  from 1 to  $N$  over  $n_i$ .



## Chapter 5

# Conclusions

We shall conclude with a reflection of what has been accomplished throughout this thesis. We have introduced error terms into models which had previously only been studied without them, or at least only spuriously included them. The situations in which we applied them relied mostly on estimators, some of which work well with error terms, while others don't.

Including error terms made the models more complicated, but also more realistic, as they simulate real life measurements, and all measurements are prone to error. In particular, the results are relevant for application, as setting the price and advertisement budget in areas of uncertainty and little information is a problem many companies face.

In Chapter 2 we introduced error terms to an allocation problem that has so far only been studied in the deterministic case. The introduction of error terms led to issues with the original proposed solution for the deterministic case, and they were remedied by the application of an exploration-exploitation algorithm. A Monte Carlo simulation was conducted to analyze the behavior of the algorithm and compare it to rules of thumb commonly used by practitioners. Within the Monte Carlo study, we varied the budget, the error terms and the functional forms, including the parameters regarding the saturation levels and the elasticities at the optimum. The results demonstrated that our proposed algorithm outperforms the rules of thumb. In particular, we recommend the application of our algorithm as opposed to the rules of thumb.

In Chapter 3 we conducted a Monte Carlo simulation study to compare eight pricing procedures in a monopolistic setting. Within this study, we varied the cost, number of iterations, error terms, the functional forms and their parameters regarding the range of the codomain. The results were rather clear, even when grouping the experimental conditions with respect to levels of cost, elasticity and disturbance level. In general, we recommend the application of the rule 'LM', which performs a regression of the data under the assumption that the graph is a straight line, and using the known formula (3.10) for the optimum in this case.

The dataset that was the result of the simulation of Chapter 2 had a property that indicated the usage of a full interaction model. That made the comparison tedious enough to warrant its own article, detailed in Chapter 4. We used the analogy between ANOVA and OLS to study the cell means within an ANOVA with the tools of the OLS setting. We started by showing a relationship between the two OLS estimators before and after a transformation that switched the reference category with a non-reference category. This was accomplished by pulling a certain matrix ‘through the OLS’. In an ANOVA context, we were able to interpret the situation as a model with all possible higher interactions, and then construct the correct null hypotheses for comparisons within that model.

This concludes the theory. The most exciting prospect of the thesis as it stands is a possible application in an actual company. For example: How does the method developed in Chapter 2 fare in practice? Once a realistic interval for the elasticities has been determined, the algorithm is ready to go and we are looking forward to practitioners using the method to increase their sales. While no *new* method for pricing was developed in Chapter 3, it is still relevant for application in a field where no or little information is available. In particular, it can be applied to the individual situation a company is in, since the variables can be chosen quite liberally. Finally, the result connecting the full ANOVA to the highest-interaction OLS and the results about the hypotheses might only be of theoretical or academical interest, however it is possible that it is a rich area of previously untapped mathematical theories that might be developed starting from these results.



# Bibliography

- Abadir K, Magnus J (2005) Matrix Algebra. Econometric Exercises, Cambridge University Press
- Aherne FJ, Thacker NA, Rockett PI (1997) The Bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika* 32(4):1–7
- Albers S (1997) Rules for the allocation of a marketing budget across products or market segments. In: Arnott (1997), pp 1–17
- Albers S (1998) Regeln für die Allokation eines Marketing-Budgets auf Produkte oder Marktsegmente. *Z betriebswirt Forsch* 50(3):211–235
- Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral ecology* 26(1):32–46
- Arnott D (ed) (1997) Marketing: Progress, Prospects, Perspectives, Proceedings of the 26th EMAC Conference, Warwick Business School
- Assmus G, Farley JU, Lehmann DR (1984) How advertising affects sales: Meta-analysis of econometric results. *J Market Res* pp 65–74
- Bates D, Watts D (1988) Nonlinear Regression Analysis and Its Applications. Wiley
- Baumol WJ, Quandt RE (1964) Rules of thumb and optimally imperfect decisions. *The American Economic Review* 54(2):23–46
- Berwin A, Turlach R (2013) quadprog: Functions to solve Quadratic Programming Problems. R package version 1.5-5
- Beswick CA, Cravens DW (1977) A multistage decision model for salesforce management. *J Market Res* 14:135–144
- Bijmolt THA, Van Heerde HJ, Pieters RGM (2005) New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research (JMR)* 42(2):141 – 156
- Camerer CF, Johnson EJ (1991) The process-performance paradox in expert judgment - how can experts know so much and predict so badly? In: Ericsson KA, Smith J (eds) *Toward a General Theory of Expertise: Prospects and Limits*, Cambridge University Press, Cambridge, pp 195–217

- Cameron AC, Trivedi PK (2005) *Microeconometrics, Methods and Applications*. Cambridge University Press
- Chiang AC (1984) *Fundamental Methods of Mathematical Economics*, 3rd edn. McGraw-Hill, Auckland
- Cohen J (1978) Partialled products are interactions; partialled powers are curve components. *Psychological Bulletin* 85(4):858 – 866
- Cohen J, Cohen P, West S, Aiken L (2013) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Taylor & Francis
- Cook R, Weisberg S (1999) *Applied Regression Including Computing and Graphics*. Wiley
- Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE (1996) On the lambert W function. *Advances in Computational Mathematics* 5(1):329–359
- East R, Wright M, Vanhuele M (2013) *Consumer Behaviour: Applications in Marketing*, 2nd edn. Sage Publications, Thousand Oaks
- Embretson SE (1996) Item response theory models and spurious interaction effects in factorial anova designs. *Applied Psychological Measurement* 20(3):201–212
- Fischer M, Albers S, Wagner N, Frie M (2011) Practice prize winner—dynamic marketing budget allocation across countries, products, and marketing activities. *Market Sci* 30(4):568–585
- Fischer M, Albers S, Wagner N (2013) Investigating the performance of budget allocation rules: A monte carlo study. *MSI reports* pp 13–114
- Gahler D, Hruschka H (2018a) Resource allocation procedures for unknown sales response functions with additive disturbances, working paper
- Gahler D, Hruschka H (2018b) Testing effects of categories of one independent variable in full anova models, as of yet unpublished
- Gigerenzer G, Gaissmaier W (2011) Heuristic decision making. *Annual Review of Psychology* 62(1):451–482
- Goldfarb D, Idnani A (1983) A numerically stable dual method for solving strictly convex quadratic programs. *Math Program* 27(1):1–33
- Greene WH (1993) *Econometric Analysis*, 2nd edn. Prentice Hall
- Gupta S, Steenburgh T (2008) Allocating marketing resources. *Marketing Mix Decisions: New Perspectives and Practices*, Roger A Kerin and Rob O'Regan, eds, American Marketing Association, Chicago, IL pp 90–105
- Hanssens DM, Parsons LJ, Schultz R (2001a) *Market Response Models*. Kluwer Academic Publishers

- Hanssens DM, Parsons LJ, Schultz RL (2001b) Market response models: econometric and time-series research. Kluwer Academic Publishers, Boston, MA
- Hirschey M, Pappas JL, Wigham D (1993) Managerial Economics. European Edition. The Dryden Press, London
- Hruschka H (1996) Marketing-Entscheidungen. Vahlen
- Huber P (1981) Robust Statistics. Wiley Series in Probability and Statistics, Wiley
- Jagpal S (1999) Marketing Strategy and Uncertainty. Oxford University Press
- JCGM-WG1 (2008) Evaluation of measurement data — guide to the expression of uncertainty in measurement. Standard, International Organization for Standardization
- Jobson JD (1991) Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design. Springer
- Kalos MH, Whitlock PA (2009) Monte carlo methods. John Wiley & Sons
- LaForge RW, Cravens DW (1985) Empirical and judgment-based sales-force decision models: A comparative analysis. *Decis Sci* 16(2):177–195
- Leeflang PS, Wieringa JE, Bijmolt TH, Pauwels KH (2015) Modeling Markets. Springer
- Malhotra NK (2007) Marketing Research - An Applied Orientation, 5th edn. Pearson Prentice Hall
- Mantrala MK, Sinha P, Zoltners AA (1992) Impact of resource allocation rules on marketing investment-level decision and profitability. *J Market Res* 29(2):162
- March JG (1991) Exploration and exploitation in organizational learning. *Organization science* 2(1):71–87
- Markovitch DG, Steckel JH, Michaut A, Philip D, Tracy WM (2015) Behavioral reasons for new product failure: Does overconfidence induce overforecasts? *Journal of Product Innovation Management* 44(5):825–841
- Maurice C, Thomas CR, Smithson CW (1992) Managerial Economics, 4th edn. Irwin, Homewood, IL.
- Monroe KB (1999) Pricing. Making Profitable Decisions, 2nd edn. Mc Graw-Hill
- Neter J, Wasserman W (1974) Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. R. D. Irwin
- Rudin W (1976) Principles of mathematical analysis, 3rd edn. McGraw-Hill Book Co., New York, international Series in Pure and Applied Mathematics
- Seber G, Wild C (1989) Nonlinear Regression. Wiley

- Sethuraman R, Tellis GJ, Briesch RA (2011) How well does advertising work? generalizations from meta-analysis of brand advertising elasticities. *J Mark Res* 48(3):457–471
- Simon H, Fassnacht M (2016) *Preismanagement*, 4th edn. Springer Gabler, Wiesbaden
- Sinha P, Zoltners AA (2001) Sales-force decision models: Insights from 25 years of implementation. *Interfaces* 31(3\_supplement):S8–S44
- Stier W (1999) *Empirische Forschungsmethoden*. Springer
- Thore S (1964) Price strategies of an 'ignorant' monopolist. In: Wold H (ed) *Econometric Model Building: Essays on the Causal Chain Approach*, North-Holland, Amsterdam, chap 7, pp 283–296
- Tukey JW (1949) Comparing individual means in the analysis of variance. *Biometrics* 5(2):99–114

# Index

- Allocation, 18, 21
- Amoroso-Robinson condition, 46
- ANOVA, 12, 56
- Balanced design matrix, 75
- Bhattacharyya coefficient, 12, 56
- Category, 102
- Competitor category, 90
- Cost function, 12
- Dimension of a square matrix, 72
- Dummy-coding, 13, 102
- Elasticity, 13
  - Arc elasticity, 13, 28, 49
  - Point elasticity, 13, 22, 46
- Exploration-Exploitation, 27
  - Exploitation, 28
  - Exploration, 27
- F-test, 13
- Forgone profit, 46, 48
- Full ANOVA model, 70
- Full model, 70
- Generic design matrix, 75
- Heuristic, 14
- Interaction, 14
- Lambert-W-function, 14, 53
- Length
  - of a design matrix, 75
  - of a manifest vector, 75
  - of a matrix, 72
- Linear model, 14
- Manifest vector, 75
- Monte Carlo, 14
- Objective function, 10, 21, 33
- OLS, 14
- Own category of a variable, 90
- Performance variable, 90
- Pricing rule, 49
  - AE, 49
  - CC, 50
  - CS, 51
  - DC, 51
  - LM, 50
  - MM, 50
  - RP, 51
  - SAE, 49
- Primary Variable, 90
- Profit Function, 48
- Reduced design matrix, 75
- Reference category, 13
- Response function, 9
  - Price response function, 46
    - Functional forms, 52
    - Stochastic response function, 48
  - Sales response function, 9, 21
- Robustness, 15, 36, 58
- Rule of Thumb, 31
- Secondary variable, 90
- Tukey-Test, 15, 56
- Usual rows, 75